Patterns of Mutation in the Human Genome

Alan James Hodgkinson

D. Phil

University of Sussex

May 2011

Table of contents

Declaration

I hearby declare that this thesis has not been, and will not be, submitted in whole or in part to any other university for the award of an other degree.

Signature……

Acknowledgments

I would like to thank my supervisor, Adam Eyre-Walker, for all his help and support over the past three years. Not only has he been an encouraging and inspiring mentor, but he has also provided a welcome distraction during many social outings. Furthermore, I would like to recognize the freedom he has given me to pursue interesting aspects of my thesis, which has enabled me to develop as a scientist. Without him, I feel this experience would have been far less enriching.

I would also like to thank my co-supervisor David Waxman for help on various mathematical issues, as well as being a source of entertainment at lunchtime and after hours. My office mates past and present – Toni, Pierre and Jess – have made coming in to work much more enjoyable; Toni in particular has been an encyclopedic reference for any computing issues. Furthermore, my numerous sporting friends – all those who play football on a Thursday and badminton on a Wednesday – have helped me to relax during the stressful periods. Similarly, I would like to thank the many wonderful people in the evolution and ecology departments and the Eyre-Walker group; Jonathan and Lynne in particular for many lively discussions.

Finally, I would like to thank my friends and family for all their support, and my girlfriend Kate for numerous encouraging chats and for understanding the late evenings and weekends.

Preface

The research presented here was carried out at the University of Sussex. Author contribution and publication details are as follows:

Chapter 2 has been published as: Hodgkinson, A., E. Ladoukakis, and A. Eyre Walker. 2009. Cryptic Variation in the Human Mutation Rate. *Plos Biology* **7**:226-232. The chapter presented here also includes some updated analyses. The study was designed by all authors. Adam Eyre-Walker (AEW) conducted the analysis on strand specific mutation rates (work contributing to appendix 2.6a), Vini Pereira analysed the gene expression data as documented in section 2.3.6, and I conducted all other analyses. AEW and I wrote the final manuscript.

Chapter 3 has been published as: Hodgkinson, A., and A. Eyre-Walker. 2010. The Genomic Distribution and Local Context of Coincident SNPs in Human and Chimpanzee. *Genome Biology and Evolution* **2**:547-557. AEW and I designed the study, I conducted all the analyses and I wrote the chapter with input from AEW.

Chapter 4 has been published as: Hodgkinson, A., and A. Eyre-Walker. 2010. Human triallelic sites: evidence for a new mutational mechanism? *Genetics* **184**:233-241. AEW and I designed the study, I conducted all the analysis and I wrote the chapter with input from AEW.

For chapter 5, I designed the study with input from AEW, I conducted all the analysis and I wrote the chapter with input from AEW.

University of Sussex

Alan James Hodgkinson

Patterns of Mutation in the Human Genome

Summary

The processes that underlie point mutations in the human genome are largely unknown. However, the cumulative effect of these processes have a large impact on how mutation rates vary across a number of different scales and contexts, and consequently guide our understanding of human disease and evolution. Although variation in the mutation rate has been characterized on many different levels, it is not fully understood the extent to which the rate of mutation can vary outside of the general patterns already observed. Beginning with the human genome project, many studies have produced large unbiased sequence datasets within a number of human populations. To that end, we analysed a number of sequence datasets in an attempt to better understand the patterns and causes of variation in the rate of mutation that exists across the genome. Firstly, we find that the mutation rates of single sites vary by more than is currently understood, and that this variation is not associated with any specific process or feature on either a local or genomic scale. Although we have been unable to uncover the source of such variation, understanding the range of mutability at sites in the human genome is important since it may point to functional regions, disease phenotypes and prompt further ideas on the underlying mechanisms associated with such a result. Furthermore, we find evidence that a mutational process that can generate the simultaneous production of two new alleles within the same individual during a single, or tightly linked series of mutation events increases the number of tri-allelic sites in the human genome. There are a number of potential mechanisms that may drive this process, and the consequences of such an event may be far reaching, as the generation of two new alleles at a single site in functional regions may allow a more rapid exploration of evolutionary space. Furthermore, this process appears to make a reasonable contribution to variation in the human genome, thus providing a substrate for evolutionary change. Finally, we observe significant variation in the mutation rate over all scales in cancer genomes. Part of this

result can be explained by the actions of specific carcinogens, however it is striking that patterns of mutation can be both consistent across different cancer types, but also very different between individuals with the same type of cancer over different scales. This result points to the idea that the patterns of mutation may vary widely between different genomes under different conditions, and the identification of general patterns in a small number of samples may not fully describe the extent to which mutation rates can vary. Taken together, these conclusions suggest that the patterns and processes underlying mutation are highly complex, and require further analysis if they are to be fully understood.

1. General Introduction

Mutation is arguably the most fundamental of all genetic processes, generating, as it does, the genetic variation that contributes to genetic disease, and upon which natural and artificial selection can act. Variation in the rate of mutation as a function of local context and genomic location has many important implications for the study of evolution and disease. Primarily an understanding of the mutability of a site or region can give insight into the rate of evolution across species via phylogenetic trees inferred from molecular changes, within populations and amongst regions of the genome. For example, distinguishing between the rates of change at synonymous and non-synonymous sites has led to an understanding of the direction of selection (Mcdonald and Kreitman 1991; Nielsen et al. 2007). Furthermore, considering the patterns of mutation is important for the identification of genes associated with disease and understanding the processes occurring at such sites (Antonarakis, Krawczak, and Cooper 2000; Kondrashov 2003; Stenson et al. 2009), as well as the identification of functional regions in the genome.

1.1 Methods

The analysis of sequence data has become an important tool in understanding mutation rates in mammalian and other genomes. While initial studies relied on limited amounts of sequence data, large sequencing projects have allowed the use of more sophisticated and robust techniques to analyze patterns of mutation. Whole genome sequences such as human (Lander et al. 2001; Venter et al. 2001), chimp (Mikkelsen et al. 2005) and

mouse (Waterston et al. 2002) have become available, and improvements in sequencing technology (for review see (Mardis 2008) have made large re-sequencing projects such as SeattleSNPs (www.pga.gs.washington.edu), the Environmental genome project (www.egp.gs.washington.edu/) and 1000 genomes project (www.1000genomes.org) more viable, generating a large amount of sequence data for analysis. Robust analysis of mutation patterns relies heavily on the correct selection of sequence data and an appropriate methodological approach (Ellegren, Smith, and Webster 2003; Baer, Miyamoto, and Denver 2007; Kondrashov and Kondrashov 2010). A common technique is to compare sequences between species and analyze divergence data (Wolfe and Sharp 1993; Hwang and Green 2004; Pink et al. 2009), however some studies focus on sequences within a species, utilizing the homologous nature of pseudogenes (Gojobori, Li, and Graur 1982; Blake, Hess, and Nicholsontuell 1992; Casane et al. 1997; Li, Yi, and Makova 2002) or repetitive sequences (Gaffney and Keightley 2005; Walser, Ponger, and Furano 2008). Furthermore, it is generally assumed that the sites being considered are neutral and hence that the pattern of substitution reflects the pattern of mutation; some analyses explicitly test for selection while others do not. Many early studies consider synonymous sites in coding sequences when estimating rates of mutation, and although synonymous sites do not lead to an alteration in the amino acid sequence, it has been questioned whether these sequences represent truly neutral sites (Gaffney and Keightley 2005; Eory, Halligan, and Keightley 2010). For this reason, researchers must be cautious when interpreting results from such data. Other studies use intronic sequences that are thought to be largely unaffected by selection (Dermitzakis, Reymond, and Antonarakis 2005; Asthana et al. 2007; Eory, Halligan, and Keightley 2010) and similar patterns of mutation to those in intronic sequence have been observed in whole genome and intergenic data (Mikkelsen et al.

2005; Taylor et al. 2006), allowing for a consideration of consistency across these sequence types.

In this introduction we consider what has been learned from sequence analysis about variation in mutation rates in the mammalian nuclear genome, and possible explanations for the patterns observed. We focus entirely on point mutations, which occur roughly 40 and 100 times more often than deletion and insertion events, respectively (Ophir and Graur 1997), and consider the patterns of mutation on a local, regional and chromosomal level. Due to the considerations above, we also assume that substitution rates are equivalent to mutation rates.

1.2 Variation in the Mutation rate

1.2.1 Sequence Context Effects

The mutation rate of single nucleotides is the smallest scale upon which variation can occur, and there is much disparity at this level. Transition mutations, which change a purine (adenine and guanine) to another purine or a pyrimidine (cytosine and thymine) to another pyrimidine, occur roughly twice as often as transversions, which change a purine to a pyrimidine or *vice versa*, with estimates of the transition/transversion ratio ranging from 1.55 to 2.5 (Gojobori, Li, and Graur 1982; Li, Wu, and Luo 1984; Blake, Hess, and Nicholsontuell 1992; Hess, Blake, and Blake 1994; Krawczak, Ball, and Cooper 1998; Zhang et al. 2007); studies with the most comprehensive datasets estimate the ratio to be ~1.9 (Zhao and Boerwinkle 2002). The excess of transitions is most

likely due to transitions only requiring one rare tautomeric form of a nucleotide to create a stable purine-pyrimidine mismatch that can avoid detection by repair enzymes, whereas transversions require two tautomeric forms to create a purine-purine mismatch (Topal and Fresco 1976). However, although transitions are generally more common than transversions, there is considerable variation in the transition-transversion ratio, particularly within mitochondria (Belle et al. 2005), with some genomes even showing no excess of transitions (Keller, Bensasson, and Nichols 2007). A bias in the direction of mutation has also been observed in many studies, with G:C→A:T mutations occurring more often than the reverse (Gojobori, Li, and Graur 1982; Li, Wu, and Luo 1984; Blake, Hess, and Nicholsontuell 1992; Hess, Blake, and Blake 1994; Hwang and Green 2004). More specifically, in the most comprehensive analysis of mutation rates, Hwang and Green (2004) estimated that the frequency of mutations in humans follow the order, C/G→T/A > T/A→C/G > C/G→A/T > G/C→C/G > A/T→C/G > T/A→A/T, which is similar to that suggested by Li, Wu and Luo (1984) and Ebersberger *et al.* (2002), who used different methods and types of sequence.

Neighbouring nucleotides also have a large effect on the mutation rate of a particular site. The most obvious example is that of transitions at CpG sites, which undergo mutation roughly 8-18 times more often than other sites in vertebrates (Coulondre et al. 1978; Bird 1980; Blake, Hess, and Nicholsontuell 1992; Nachman and Crowell 2000; Hwang and Green 2004; Elango et al. 2008). CpG dinucleotides are generally methylated in mammals and since methyl-cytosine is unstable it undergoes high rates of spontaneous deamination to thymine, which is repaired less efficiently than the usual product of cytosine transitions, uracil (Coulondre et al. 1978). CpG transversions also mutate at a higher rate than other transversions (Cooper and Krawczak 1990; Blake,

Hess, and Nicholsontuell 1992; Nachman and Crowell 2000; Hwang and Green 2004), and it is thought that this might be due to error prone repair at altered guanines after methylation and deamination at cytosines (Blake, Hess, and Nicholsontuell 1992).

It has been suggested that mutation rates can be broadly divided into four groups in humans: CpG transitions > CpG transversions ~ non-CpG transitions > non-CpG transversions (Zhang et al. 2007). However, within each of the non-CpG groups there is variation in the rate of mutation (Hwang and Green 2004; Zhang et al. 2007), with Hwang and Green (2004) identifying 14 types of substitution that show similar rates across the mammalian phylogenetic tree. The mutation rate associated with neighbouring nucleotides is estimated to vary over 52 to 72-fold (Hess, Blake, and Blake 1994; Hwang and Green 2004), which reduces to 10-fold when CpG transitions are excluded (Hwang and Green 2004). More specifically, mutations appear to be more frequent in the presence of certain flanking nucleotides, with C>A>G>T being the order of prevalence at the 5' flanking site (Zhao and Boerwinkle 2002); this is likely to be the most accurate ordering since Zhao and Boerwinkle (2002) performed the analysis using SNPs across the whole human genome, whereas contradictory predictions by Blake, Hess, and Nicholsontuell (1992) were generated using pseudogenes with a GC content of ~57%, which is much higher than the genome average of 41%. One specific feature of these patterns is that there is an increased rate of mutation at alternating purines and pyrimidines (Blake, Hess, and Nicholsontuell 1992; Hess, Blake, and Blake 1994) – an effect that is also seen at CpG nucleotides with more mutable doublets being flanked by a 5' pyrimidine and a 3' purine (Krawczak, Ball, and Cooper 1998; Zhao and Boerwinkle 2002) – that is likely due to the more stable nature of alternating purines-

pyrimidines, which form inter-base-pair hydrogen bonds resulting in mismatches being less detectable by repair machinery (Kennard and Hunter 1991).

Finally, the effects of local context can be important beyond the neighbouring nucleotides. Although it is thought that the impact of the surrounding sequence quickly diminishes beyond the flanking nucleotides, and possibly the next two or three positions (Krawczak, Ball, and Cooper 1998), it has been shown that there is significant heterogeneity in nucleotide composition up to as far as 200bp away for all classes of mutation (Zhao and Boerwinkle 2002; Elango et al. 2008). One particularly well-studied example is that of CpG dinucleotides, where an increase in AT content around CpGs results in a higher rate of mutation (Fryxell and Zuckerkandl 2000; Fryxell and Moon 2005; Elango et al. 2008). Elango *et al.* (2008) have shown that this context dependence decays in an exponential fashion over about 1500bp either side of the CpG. It seems likely that the increased mutability of CpGs in AT rich sequences is caused by a process suggested by Fryxell and Zuckerkandl (2000) in which individual GC nucleotide pairs remain paired for ~3 times longer than AT pairs during DNA 'breathing' that occurs regularly in the cell (Leroy et al. 1988). Cytosine deamination occurs ~143 times more often on ssDNA than it does on dsDNA (Frederico, Kunkel, and Shaw 1990), thus the mutability of CpGs is closely linked to the melting temperature of the surrounding DNA. A 10% decrease in GC content reduces the melting temperature of a sequence by 4.1$_{o}$C, increasing the deamination of methylated cytosine by 2-fold (Fryxell and Zuckerkandl 2000). It is unknown whether similar processes could have an impact on other types of mutation, although it seems unlikely as most other mutational classes seem to be a result of replication errors, rather than

methylation and deamination (see later).

Whereas the CpG effect is well understood, the reason for other context effects is not so evident. There are two types of general feature that may impact upon the rate and type of mutation in different contexts: exposure to exogenous factors and consequences of endogenous error prone processes such as replication and repair (Cooper and Krawczak 1990). Examples of exogenous factors include UV light, which increases the rate of mutation at pyrimidine dimers (Pfeifer, You, and Besaratinia 2005), and chemical mutagens such as those found in tobacco smoke, which increase the rate of C>A mutations (Pfeifer et al. 2002). More generally, it appears that guanine is the most frequently attacked base, being preferentially alkylated by many chemicals (for review on this and many other mutation processes, see (Boulikas 1992). However, although exogenous factors play a part in increasing and biasing patterns of somatic mutation, it seems unlikely that they have a large effect on germline mutations, since it is thought that most germ-line mutations are a consequence of DNA replication. The evidence for this comes from several sources. First, the Y-chromosome appears to have a higher mutation rate than the autosomes, which have higher mutation rates than the X-chromosome (Ebersberger et al. 2002; Malcom, Wyckoff, and Lahn 2003; Mikkelsen et al. 2005; Goetting-Minesky and Makova 2006); this is consistent with most mutations occurring through DNA replication, since more germ-line replications occur in the male germ-line, than in the female, and the Y-chromosome spends all of its time in the male germ-line (Miyata et al. 1987) (see discussion below). Second, Cooper and Krawczak (1990) have noted that the pattern of mutation exhibited *in vitro* by the two main replication enzymes in mammals, DNA polymerase α and β, are similar to those observed in human disease mutations when proof-reading is no longer active. This

implies that a substantial proportion of point mutations are caused by mis-incorporation of nucleotides during DNA replication.

During the processes of replication and repair, there are many examples of physical and enzymatic processes that can bias the rate of mutation. Transient misalignment is a process by which a single nucleotide (usually one that is part of repetitive DNA) can loop out from the double helix during misalignment, with the following mis-incorporation and re-alignment of sequences resulting in a mismatch (Kunkel 1985). Indeed, some studies have reported an excess of mutations in which the newly inserted base is identical to one of the flanking nucleotides, a pattern consistent with transient misalignment process (Cooper and Krawczak 1990; Todorova and Danieli 1997; Krawczak, Ball, and Cooper 1998). Similarly, the stability of local DNA context can vary (SantaLucia, Allawi, and Seneviratne 1996) and the mutability of particular nucleotide combinations significantly co-varies with this feature (Cooper and Krawczak 1990; Krawczak, Ball, and Cooper 1998). Furthermore, some mutations are more likely to be read through during replication, allowing the error to be reproduced in subsequent DNA. For example, G.T mispairs can be accommodated with minimal distortion of the DNA helix, as many functional groups of the nucleotides can bind to the opposite base in much the same way as Watson-Crick pairings (Kennard and Hunter 1991). It is also apparent that some repair enzymes show a bias in the direction of repair dependent on mutation type and local context (Brown and Jiricny 1988). Similarly, adenine is preferentially inserted at apurinic sites (Loeb 1985). Although this is not a comprehensive list of processes that bias mutation, it does give an indication of the types of process that can affect mutation rates on a local scale. It seems likely that a combination of these factors leads to variation in the mutation rate on this scale and

ultimately it is difficult to assess whether one process has more impact than any other by analyzing sequence data alone.

Although the rates of mutation associated with particular types of mutation and their local context have been well characterized, it is unknown the extent to which these factors govern the overall mutability of a site. It seems likely that there is variation in the mutation rate within a particular class of mutation and within a particular context. Other factors may impact on the mutability of a site outside of local primary sequence context, resulting in far more variation in the mutation rate of single sites than is currently understood. Although it is difficult to identify general processes that may alter the mutability of particular sites, in chapter 2 we attempt to quantify the variation in the mutation rate that cannot be explained by known factors. In this way, we can better understand the pattern and frequency of change in the human genome, for example the length of conserved regions we might expect to see by chance, or impact of hypermutation in disease phenotypes. Understanding the variation in the rate of mutation is also important in evolutionary inference, particularly in the reconstruction of ancestral sites and phylogeny. Furthermore, understanding variation in the mutation rate of single sites may lead to a greater understanding of why some single sites are more prone to mutation than others. In chapter 3 we consider the spatial distribution of sites we suspect are undergoing higher rates of mutation, in attempt to answer such questions.

1.2.2 Other mechanisms

Some features of mammalian genomes imply that there are more complex processes

outside of the general effects mentioned above that lead to variation in the mutation rate. There are many types of mutation hotspot that are associated with specific motifs, mostly driven by association with particular molecules (Rogozin and Pavlov 2003). For example, Todorova and Danieli (1997) showed there to be a large excess of mutations at a specific motif, TG[A/G][A/G][G/T][A/C], associated with α-polymerase pause sites in the human dystrophin gene where the arrest of α-polymerase during the replication cycle is thought to reduce its accuracy of incorporation or repair, and this has also been observed in the human lioprotein lipase gene (Templeton et al. 2000).

We have so far only considered single point mutations, however, it is has been shown that mutations can sometimes occur simultaneously at adjacent sites. This was first elegantly demonstrated by Averof *et al.* (2000), who showed that the rate of substitution between TCN and AGY serine codons (where N is any nucleotide and Y is a pyrimidine) was faster than one would expect given the rate of single mutation. Averof *et al.* (2000) estimated the rate of doublet mutation to be ~2% of the rate of single nucleotide changes, averaging across diverse taxomonic groups. In contrast, in mammals the rate has been estimated as 0 (i.e. no excess at all) (Silva and Kondrashov 2002) and 0.1%-0.44% (Kondrashov 2003; Smith, Webster, and Ellegern 2003). The rate of mutations occurring at adjacent sites is also considered in chapter 4 of this thesis. Doublet mutations may have a great impact if they occur in coding regions, as the number of steps required to move between different amino acids could be dramatically reduced via this process.

It has recently been suggested that a similar process may also occur across DNA strands, with an error at one site increasing the rate of mutation at the site opposite

(Walser, Ponger, and Furano 2008). Such a mechanism could explain why the rates of CpG and non-CpG substitution are correlated across genes (Walser, Ponger, and Furano 2008); a C>T mutation at CpG sites may be preferentially preserved by DNA polymerase iota and the subsequent G/T mismatch may then be subject to error prone repair, leading to a mutation that would no longer be seen as a CpG site in sequence analysis. It is unknown how frequent such a process might be and whether it could have far reaching impacts on the human genome in generating variation and thus providing a medium for evolutionary change. In chapter 4, I consider whether a mechanism that generates two new alleles simultaneously can explain why there is an excess of tri-allelic SNP sites in the human genome, thus attempting to identify whether such a process can generate appreciable genetic variation.

1.2.3 Regional patterns of mutation

Neighbouring nucleotide effects generate variation in the mutation rate on a single nucleotide scale. However, it is evident that the mutation rate also varies at larger scales. At a small regional scale it has been shown that insertion/deletion events (indels) increase the rate of mutation in adjacent sequences, with the largest effects observed up to ~50bp from the indel, but reaching as far as ~400bp away. This effect is thought to be caused by problems in the pairing of chromosomes in individuals heterozygous for an indel (Tian et al. 2008).

Furthermore, it has been proposed that recombination itself may be mutagenic (Lercher and Hurst 2002; Hellmann et al. 2003; Hellmann et al. 2005), as primate and rodent divergence data correlates with recombination rates, and this may result in small-scale

regional variation in mutation rates. Hellmann *et al.* (2005) also showed that nucleotide diversity within chimpanzee does not correlate with human recombination rates, which can be explained by the two species differing dramatically in patterns of recombination (Winckler et al. 2005); if the patterns were driven by selection one might expect human recombination rates to correlate with chimpanzee SNPs since the two species are almost identical in the locations of coding regions (Mikkelsen et al. 2005). It is also unlikely that the process is driven entirely by biased gene conversion, a process in which mismatches are preferentially repaired to GC over AT around double strand breaks at recombination events (Marais 2003), as GC conserving and GC changing mutations both correlate with recombination (Lercher and Hurst 2002). There is also some debate as to the strength of fixation bias in the direction of GC around recombination events, and some studies may have overestimated this effect as a consequence of mis-inferring the ancestral allele at a mutation event (Hernandez et al. 2007).

Mutation rates also vary within chromosomes on a much larger regional scale. The evidence for this comes from a number of mammalian species and sequence types. The earliest confirmation came from the demonstration that the rate of synonymous substitution (K_s) varied significantly across the genome and that it was correlated to GC content (Wolfe, Sharp, and Li 1989; Wolfe and Sharp 1993). Subsequently, Matassi, Sharp, and Gautier (1999) showed that the differences in K_s between genes that were located close to each other on the human genome were significantly lower than values observed at random gene pairs using human-mouse divergence data. The same pattern is also seen in rodents (Lercher, Williams, and Hurst 2001; Williams and Hurst 2002). Although these results must be treated with caution, since there is evidence that selection can act upon synonymous sites in mammals (Chamary, Parmley, and Hurst

2006), subsequent analyses using non-coding sequences have confirmed that there is large scale variation in the mutation rate (Casane et al. 1997; Smith, Webster, and Ellegren 2002; Mikkelsen et al. 2005; Zhang et al. 2007), even when sequences likely to be subject to selection have been removed.

The most extensive analysis of large-scale variation in the mutation rate has been performed by Gaffney and Keightley (2005). They considered the scale over which the mutation rate varies by analyzing the spatial autocorrelation in the divergence between ancestral repeats found in mouse and rat. They showed that the mutation rate varies little within 100kb, but that it then decays exponentially until there is little correlation between adjacent blocks at a scale of 10-15MB. They further show that all the autocorrelation above 1MB can be explained in terms of the correlation between 1MB blocks, suggesting that the scale over which the mutation rate varies is less than 1MB but greater than 100KB. Other studies have argued that variation within chromosomes can be found to act over many scales; blocks of 50kb show significant variation (Elango et al. 2008) and a non-random distribution of mutations have been observed in regions of between 1kb and 10kb (Silva and Kondrashov 2002). However, it seems that the strongest effect within chromosomes occurs on a scale of 1MB (Silva and Kondrashov 2002; Gaffney and Keightley 2005; Mikkelsen et al. 2005).

Whatever is generating variation in the mutation rate over a scale of 1MB must be a factor that operates on this scale. There are a number of possible candidates. It has recently been shown that the rate of mutation is correlated to replication time (Stamatoyannopoulos et al. 2009) with late replicating regions of the genome having a mutation rate that is ~30% higher than early replicating regions. This increase is

observed for both CpG and non-CpG mutations, although it is much more evident for the former. Stamatoyannopoulos *et al.* (2009) suggest that this is due to accumulation of single stranded DNA late in the cell cycle, driven by stalling of DNA polymerase enzymes in late replicating regions due to heterochromatic DNA or by a full depletion of nucleotides pools. Alternatively, there may be more mutations occurring in late replicating regions due to collisions between replication and transcription machinery (Mirkin and Mirkin 2007), however this seems unlikely since there is a negative correlation between the two events (Chen et al. 2010). It has also been suggested that the activity and fidelity of repair enzymes may decrease throughout replication (Holmquist and Filipski 1994; Chen et al. 2010), although there is currently no evidence for this.

The idea that replication time might influence the rate and pattern of mutation has a much older history; Wolfe, Sharp, and Li (1989) originally suggested that changes in the dNTP pools through the cell cycle might induce changes in the GC content and mutation rate based upon a correlation between GC content and the rate of synonymous substitution. DNA precursor pools are very small, often only sufficient for only a few minutes of replication (Meuth 1989), thus small changes in nucleotide ratios may have a big impact on mutation rate. Unfortunately, the direction and strength of the relationship between GC content and mutation rate is open to much debate, and different studies have shown positive (Matassi, Sharp, and Gautier 1999; Smith, Webster, and Ellegren 2002), negative (Filipski 1988; Hellmann et al. 2005), and other more complex (Wolfe, Sharp, and Li 1989; Wolfe and Sharp 1993; Eory, Halligan, and Keightley 2010) correlations. However, although there is some debate about the nature of the relationship between the mutation rate and GC content, there is a consensus that

GC content explains relatively little of the variance in the mutation rate, and that there is significant variation in the mutation rate when GC content is controlled for (Matassi, Sharp, and Gautier 1999; Hellmann et al. 2005). For example, Smith, Webster, and Ellegren (2002) showed that GC content only explained 10% of the variation in substitution rates, and so other factors are clearly involved. It is perhaps surprising that mutation rates are not more strongly correlated to GC content since there is a general bias towards GC>AT mutations, however the variance in GC content across human chromosomes is not that large; 75% of 1MB regions of the genome lie within ~36-46% GC content (Venter et al. 2001). It is worth noting that the mutation rate still correlates significantly to the timing of DNA replication when GC content is controlled for (Chen et al. 2010).

Another feature that correlates with mutation rates is chromatin structure, with more mutations occurring in closed chromatin rather than open chromatin (Prendergast et al. 2007; Ying et al. 2010). This may be explained by varying repair efficiencies of different enzymes in closed and open chromatin (Filipski 1988; Sueoka 1992; Matassi, Sharp, and Gautier 1999; Gaffney and Keightley 2005). However, some doubt over a direct causal relationship between the two factors has arisen since Chen *et al.* (2010) showed that the correlation is no longer significant when controlling for replication timing. There are also various anomalies, such as the high rate of substitution on chromosome 19 despite the chromosome being particularly enriched with open chromatin (Prendergast et al. 2007), which would need to be explained under this model.

Although it is understood that the rate of mutation varies on a regional scale, it is not

fully understood which mechanism is driving the affect and whether other currently undiscovered factors may also play a part. It seems that replication timing has some effect on regional variation in the mutation rate, however since the level of variation on this scale is small compared to more local effects, these processes may ultimately have little effect on mutation rates genome wide. It is also not known how regional patterns of mutation might vary among different individuals and between different cell types, and other factors such as chromatin structure may become more important in different scenarios. In chapter 5 we consider the regional patterns of mutation in four cancer genomes in an attempt to understand whether this level of variation in the mutation rate becomes more important in disease phenotypes. We also consider whether regional variation in the mutation rate can play an important role as we consider the extent to which it might be driving changes in disease genes themselves.

1.2.4 Chromosomal variation in the mutation rate

Mutation rates are also known to vary between chromosomes. The strongest differences exist between the sex chromosomes and the autosomes, with the Y-chromosome having a higher mutation rate than the autosomes, which has a higher mutation rate than the X-chromosome. This pattern has been shown in primates (Ebersberger et al. 2002; Malcom, Wyckoff, and Lahn 2003; Mikkelsen et al. 2005; Goetting-Minesky and Makova 2006), rodents (Wolfe and Sharp 1993; McVean and Hurst 1997; Lercher and Hurst 2002; Malcom, Wyckoff, and Lahn 2003; Gaffney and Keightley 2005) and the perissodactyls (horses and rhinos) (Goetting-Minesky and Makova 2006).

Two hypotheses have been proposed to explain this pattern. First, It has been suggested

that it is due to males having a higher mutation rate than females because they undergo more cell divisions in their germ-line (Haldane 1947). It has also been proposed that the male-driven bias is a consequence of mature sperm lacking in certain repair enzymes that leads to a higher rate of mutation (Boulikas 1992) or varying levels of methylation during the perigametic interval causing higher levels of mutation in males (Russell 1999), however a dependency on the number of cell divisions is by far the most supported theory. Since the Y-chromosome is only transmitted through the male germ-line, whereas an autosome spends 1/2 its time and the X-chromosome 1/3 of its time in the male, we would expect the Y-chromosome to have a higher mutation rate than the autosomes, which in turn should have a higher mutation rate than the X, if most mutations are generated by DNA replication (Miyata et al. 1987). Several lines of evidence support this model. First, the male-to-female mutation rate, α, has been estimated from the divergence of the Y, Z and autosomes to be between ~2 and ~7 in primates (Bohossian, Skaletsky, and Page 2000; Makova and Li 2002; Mikkelsen et al. 2005; Goetting-Minesky and Makova 2006; Taylor et al. 2006), with the most reliable estimate in the 6-7 range (Taylor et al. 2006), and ~2 in rodents (Chang et al. 1994; Makova, Yang, and Chiaromonte 2004). These predictions are reasonably consistent with estimates of the ratio of the numbers of male to female germ-line cell divisions; in humans it has been estimated that males have undergone ~6 times as many germ-line as females when they are 20 years old, ~10 times as many at 25 years and ~27 times as many at 40 years (Crow 2000; Li, Yi, and Makova 2002). In contrast, in mice it is estimated that the ratio of male to female of cell divisions is much lower than in humans at about 2 (Chang et al. 1994). It is worth noting that the very lowest estimates of α in primates can be explained by methodological problems. Predictions of α may vary when using limited datasets focused in small areas of the genome as mutation rates are

known to vary on a regional scale (McVean 2000; Li, Yi, and Makova 2002). Indeed, Goetting-Minesky and Makova (2006) showed that estimates of α can vary over 16-fold using different genes within the same genome; part of this large variation in the range of predicted α values can be explained by sampling bias and part by regional variation in the mutation rate. Similarly, the lowest prediction of α was obtained by a comparison between humans and chimpanzees (Bohossian, Skaletsky, and Page 2000), and here ancestral polymorphism becomes a problem since the Y-chromosome typically has very little diversity (Li, Yi, and Makova 2002). Indeed, when levels of polymorphism are accounted for in the common ancestor of humans and chimps, α predictions are much more consistent across the chromosomal classes (Ebersberger et al. 2002) and estimates of α are much more reliable when using primate data between more distantly related species, probably due to the smaller influence of ancestral polymorphisms over a longer period of time (Makova and Li 2002).

Second, it has been shown that there is much more variation in the non-CpG mutation rate between the sex chromosomes and the autosomes, than for CpG mutations (Taylor et al. 2006). This is consistent with most non-CpG mutations being generated by DNA replication, while most CpG mutations are induced by deamination, a replication independent process. Nevertheless, CpG sites show a male mutation bias of ~2-3 in humans (Mikkelsen et al. 2005; Taylor et al. 2006). This could be due to replication through a mismatch at a CpG site before it has been repaired, or due to higher levels of methylation on the X chromosome compared to the autosomes, leading to a depletion of CpG dinucleotides on the X chromosome and thus a lower CpG transition rate (Krawczak, Ball, and Cooper 1998; McVean 2000).

Third, it has been shown in birds that the Z chromosome has a higher substitution rate than the autosomes, which have a higher substitution rate than the W chromosome; in birds the female is the heterogametic sex. This is expected since the Z chromosome spends 2/3rds of its life being transmitted through the male germ-line, whereas the W chromosome is always transmitted through the female germ-line (Ellegren and Fridolfsson 1997).

The second theory to explain the differences between the sex chromosomes and the autosomes stems from an observation that estimates of α are generally larger when they are estimated from comparisons of the X and autosomes, rather than the Y and autosomes (McVean and Hurst 1997); the X-chromosome appears to have a mutation rate that is lower than expected from the male mutation bias hypothesis. It has been suggested that this might be due to strong selection to reduce the mutation rate on the X-chromosome since mutations are hemizygous, and so effectively dominant, in males and often effectively hemizygous in females because of dosage compensation. However, the theory has since been questioned by Malcom, Wyckoff and Lahn (2003), who showed that the substitution rate of the X chromosome is actually compatible with the male-driven mutation theory using a larger dataset.

As well as the differences between sex chromosomes and the autosomes, there is also significant variation in the substitution rate amongst the autosomes (Ebersberger et al. 2002; Lercher and Hurst 2002; Malcom, Wyckoff, and Lahn 2003; Gaffney and Keightley 2005; Mikkelsen et al. 2005), which implies that chromosomal differences are not entirely driven by male-driven mutation. However, the variance between chromosomes is ~10-fold smaller than the variance between sub-chromosomal sections

at the 1MB scale (Wolfe and Sharp 1993). The reasons for the between chromosomal differences remain obscure. There is a strong positive correlation in rodents between the amount of rearrangement a chromosome has undergone and the rate of substitution (Pink et al. 2009); rearrangement explains 56% of the variance in chromosomal substitution rates. This may be driven by a link between recombination and chromosomal rearrangements; recombination is thought to be mutagenic (Lercher and Hurst 2002; Hellmann et al. 2003; Hellmann et al. 2005), and thus chromosomes that undergo higher rates of recombination may contain more mutations and also contain more rearrangement events (Pink et al. 2009). Furthermore, Pink and Hurst (2010) showed that substitution rates in rodents can partially be explained by differences in average replication timing between chromosomes; however this explains little of the variance as there is only about 4.5% difference in the rate of substitution between the earliest and latest replicating autosomes. However, replication timing and chromosomal rearrangements make independently significant contributions to chromosomal mutation rates, and together it is estimated that they explain ~70% of the variance between chromosomes (Pink and Hurst 2010).

Variation in the mutation rate between the sex chromosomes and the autosomes is a well-characterized phenomenon, and there is a lot of evidence to suggest that it is driven by male mutation bias. However, the processes driving variation in the mutation rate among the autosomes are less clear and it is not certain whether patterns on this level are driven entirely by variation on a regional scale. It is difficult to see why one autosome should have a higher mutation rate than any other, outside of processes occurring on smaller scales. As with variation on a regional scale, it is also unknown whether chromosomal mutation rates might vary more widely under different conditions

and in different individuals. Again, in chapter 5 we consider the rates of chromosomal mutation in cancer genomes in order to identify whether there is variation beyond that observed in the germ line.

1.3 Concluding remarks

There is significant variation in mutation rates across various scales in mammalian genomes, and analysis of sequence data has been an important tool in revealing and explaining those patterns. However, there are many mutation patterns and molecular processes yet to be fully understood and further analysis may need to go beyond primary sequence context to higher orders of DNA conformation. The ultimate causes of mutation rate variation appear complex and no single mechanism can explain the differences that exist over many scales. However, it is interesting to consider how variation in the mutation rate is tolerated within a genome, and whether certain areas may have evolved to mutate at higher rates to increase the rate of evolution or lower rates to avoid harmful mutations (McVean and Hurst 1997; Ellegren, Smith, and Webster 2003). Perhaps the biggest mystery is whether the patterns and processes outlined above can account for the majority, if not all, of the variation in the rate of mutation that occurs across the human genome, and indeed within the human population.

In this thesis I consider other factors that can influence the rate of mutation, particularly at single sites in order to better understand how mutation rates can vary. In chapters 2 and 3, I consider whether the rate of mutation at single sites can vary independently of patterns associated with neighbouring nucleotides and other local contexts, and whether

any regional genomic feature can impact upon this. In this way we may be able to make predictions on the impact of varying rates of mutation at single sites, which can affect our outlook on evolution and disease. In chapter 4, I identify and seek to explain an excess of tri-allelic sites in the human genome by invoking a process by which two new alleles are generated at a single site in a single or tightly linked series of events. Finally, in chapter 5, I look at how patterns of somatic mutations vary over a number of different scales in cancer genomes, in an attempt to better understand how the rate of mutation can vary under different conditions and in different individuals.

2. Cryptic Variation in the Human Mutation Rate

2.1 Abstract

The mutation rate is known to vary between adjacent sites within the human genome as a consequence of context, the most well studied example being the influence of CpG dinucelotides. Here we investigate whether there is additional variation by testing whether there is an excess of sites at which both humans and chimpanzees have a single nucleotide polymorphism (SNP). We find a highly significant excess of such sites and demonstrate that this excess is not due to neighbouring nucleotide effects or natural selection, and although we are unable to rule out a contribution from ancestral polymorphism and substitutions in paralogous sequences, it seems unlikely that these processes can explain much of the excess of coincident SNPs. We therefore infer that there is cryptic variation in the mutation rate. However, although this variation in the mutation rate is not associated with the adjacent nucleotides, we show that there are highly non-random patterns of nucleotides that extend ~80bp on either side of sites with coincident SNPs, suggesting that there are extensive and complex context effects. Finally, we estimate the level of variation needed to produce the excess of coincident SNPs and show that there is possibly at least as much variation in the mutation rate associated with this cryptic process as there is associated with adjacent nucleotides, including the CpG effect. We conclude that there is substantial variation in the mutation rate that has, until now, been hidden from view.

2.2 Introduction

The mutation rate is thought to vary across the human genome on several different scales. At the chromosomal level, the Y-chromosome evolves faster than the autosomes, which evolve faster than the X-chromosome (Miyata et al. 1987; Li, Yi, and Makova 2002). This is thought to be due to males having a higher mutation rate than females. The autosomes also appear to differ in their rates of mutation for reasons that are unclear (Lercher, Williams, and Hurst 2001; Gaffney and Keightley 2005). At the next level down, there appears to be variation in the mutation rate over a scale of several hundred kilobases (Matassi, Sharp, and Gautier 1999; Gaffney and Keightley 2005), another pattern that remains unexplained. However, the most dramatic variation in the mutation rate is observed over fine scales in which adjacent sites can have very different mutation rates. In the nuclear genome, this variation has been shown to be associated with context, the best-known example being the CpG dinucleotide in mammals. CpG dinucleotides are generally methylated in mammals and since methyl-cytosine is unstable, this leads to a high rate of C->T and G->A transitions at these sites, which is about ten- to twenty-fold higher than at other sites (Coulondre et al. 1978; Bird 1980). However, the CpG effect is not the only source of fine-scale variation in the mutation rate; the rate of mutation appears to vary by about two or three-fold as a function of other adjacent nucleotides (Blake, Hess, and Nicholsontuell 1992; Rogozin and Pavlov 2003; Zhao et al. 2003; Hwang and Green 2004). Furthermore, the pattern of mutation is known to differ between the two DNA strands due to repair biases that occur during transcription, leading to an excess of G/T over A/C on the coding strand of genes (Green et al. 2003; Gibbs et al. 2004).

Although variation in the mutation rate has been well characterised in terms of adjacent nucleotides (Blake, Hess, and Nicholsontuell 1992; Zhao et al. 2003; Hwang and Green 2004), it is possible that there is other variation in the mutation rate that is associated with either distant or complex context effects, which has hitherto escaped detection. Here we investigate this question by testing whether human and chimpanzee single nucleotide polymorphisms (SNP) occur at orthologous sites in the genome. If there is variation in the mutation rate we expect to see an excess of sites at which both humans and chimpanzees have a SNP.

2.3 Materials and methods

2.3.1 Data

In order to consider the number of human and chimpanzee coincident SNPs we obtain polymorphism data from dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), which is a public repository for genetic variation. It includes variation data for single nucleotide polymorphisms (SNPs), single base insertions and deletions, multi-base small scale insertions and deletions, retro-transposable element insertions and microsattelite repeats (Sherry et al. 2001). In this study we focus on SNPs. For humans, variant entries are submitted from the SNP consortium, the genomic sequence from the human genome project, various large-scale variant detection studies and individual lab contributions, obtained through a mixture of sequencing and other genotyping methods. Chimpanzee SNPs are largely contributed by the chimpanzee genome sequencing project, but also from other individual lab contributions (Kitts and Sherry 2010). Most SNPs included in

the database are from direct submission information, however some SNP entries are a result of computational calling based on originally submitted data, where new submissions match to the reference sequence and indicate a new variant. Each SNP entry contains flanking sequence, which must be of a minimum length to increase the likelihood of a unique match to the reference sequence, population and sample sizes, a validation status, details on the detection method and may also include allele and genotype frequencies. The genomic location of each SNP is given, which is obtained by BLASTing new submissions to the reference genome. Repeats are masked during the BLAST process via the 'Dust' option and the locations of variants are only confirmed if there is a reasonable match to unmasked sequence; if this criteria is met, alignments are allowed to extend into masked regions (Kitts and Sherry 2010). Confirmed SNP entries are then linked to numerous other databases, which include PubMed (http://www.ncbi.nlm.nih.gov/pubmed) and the Online Mendelian Inheritance in Man website (http://www.ncbi.nlm.nih.gov/omim). There has been much debate on the quality of data found in dbSNP, and some researchers have highlighted errors in sequencing and SNP calling. Mitchell *et al* (2004) have suggested that ~15-17% of entries in dbSNP are sequencing errors, the majority of which arise from errors made by base calling software, particularly at heterozygous sites. Furthermore, Musumeci *et al* (2010) have shown that ~8.32% of human SNP entries may be false positives due to substitutions in paralogous sequences being incorrectly called as SNPs.

2.3.2 Coincident SNPs

We downloaded human and chimpanzee SNPs from dbSNP build 126. Dividing the data into chromosomes we BLASTed each chimpanzee SNP, along with 50bp of

flanking DNA on either side of the SNP, against a database of human SNPs. We set the BLAST parameters as follows; e-value = $1x10^{-30}$, mismatch score = -1 and simple sequence filter off. We retained those alignments which were 101bp in length, and in which the human or chimpanzee sequence showed identity at 96 sites if the SNPs were coincident, or 94 sites if they were not coincident. We adjusted the number of matches required to control for the fact that if the SNPs are not coincident then there must be two extra mismatches. We randomly chose one alignment if a chimpanzee SNP matched more than one human SNP at the levels of identity we set; we obtained very similar results removing these cases from the analysis. The alignments were trimmed to 40bp either side of the central chimpanzee SNP because there is a slight bias away from finding human SNPs at the edges of the chimpanzee query sequence. This bias occurs because SNPs, being classified as mismatches, tend to cause BLAST to prematurely terminate the alignment. To perform the analysis of triplet frequencies we downloaded an extended flanking sequence for the chimpanzee SNPs analysed.

The macaque SNPs were kindly provided by Dr Ripan Malhi (2007). We repeated the analysis as we did for chimpanzee but we relaxed the criteria used to identify orthologous human sequences containing SNPs to 86 matches if there was a coincident SNP, and 84 if there was not, with the e-value adjusted to allow this level of similarity to be found.

Sites were designated as CpG if the site, or any of the SNPs at the site, would yield a CpG dinucleotide.

2.3.3 Estimating the expected number of coincident SNPs

To estimate the expected number of coincident SNPs under the null hypothesis, ignoring any context effects, we divided the total number of alignments that contained both a human and chimpanzee SNP by 81. Since the chimpanzee SNP always occurs in the central position of the 81bp sequence, if the human SNP is distributed randomly across the alignment, it is expected to be in the central position once in every 81 alignments at random.

We estimated the expected number of coincident SNPs, taking into account the effects of adjacent nucleotides on the rate of mutation, what we term "simple" context effects, as follows. Our data consists of a set of alignments in which we have both a human and a chimpanzee SNP. We start by tabulating the numbers of each triplet, n_{xyz}, where x, y and z can be T, C, A or G, in the chimpanzee sequence in the alignments, along with the number of chimp triplets that have a human SNP opposite the central nucleotide, $n_{xyz.Hsnp}$. From these, we can estimate the probability of observing a human SNP opposite a chimpanzee triplet in our alignments: $p_{xyz} = n_{xyz.Hsnp} / n_{xyz}$. We can also calculate the frequency of each triplet in the chimpanzee sequences: $f_{xyz} = n_{xyz} / \Sigma n_{xyz}$. To calculate the probability that the human and chimpanzee SNPs are coincident, we need to take into account that there are two chimpanzee alleles, and the triplets they are a part of will have different probabilities of having a human SNP opposite them. If we knew the relative frequencies of the chimpanzee alleles we could calculate the chance of a coincident SNP as $g_y\, p_{xyz} + (1- g_y)\, p_{xy'z}$ where y and y' are the two chimpanzee alleles and g_y is the frequency of the y allele. However, we do not have allele frequency information, so we calculated the relative probabilities of each of the two ancestral

states for the chimpanzee SNP, since the ancestral allele is likely to be at higher frequency in the population. For example, let us imagine we have a CYC SNP – i.e. a Y SNP surrounded by C on both sides. The ancestral triplet could have been CCC or CTC. The probability that it was CCC can be estimated as $m_{CCC} = f_{CCC}\, r_{CCC.snp} / (f_{CCC}\, r_{CCC.snp} + f_{CTC}\, r_{CTC})$ where r_{xyz} is the rate at which triplet XYZ generates a SNP in the central position of the triplet. We estimate r_{xyz} by orienting the chimp SNPs using the human sequence, excluding coincident SNPs and SNPs for which the human nucleotide is different to both chimp alleles; let $s_{xyz.Csnp}$ be the number of chimp triplets that are inferred to have generated a SNP, then $r_{xyz.snp} = s_{xyz,Csnp} / n_{xyz}$. The expected number of coincident SNPs in each alignment is then, using the above example, $(m_{CCC}\, p_{CCC} + m_{CTC}\, p_{CTC}) / \Sigma\, p_{xyz}$, where the summation is across all the triplets in the alignment. The total number of expected coincident SNPs was simply the sum across alignments.

We used two methods to calculate the standard error for the ratio of the observed number of coincident SNPs over the expected number; we bootstrapped the data by alignment and then summed the observed and expected values across the bootstrapped datasets. However, it turned out that this was very closely approximated by assuming that the observed number of coincident SNPs was Poisson distributed and the expected value was known with no error; these are the standard errors we present.

2.3.4 Simulations

We performed a number of simulations to check that the BLAST analysis was not biased and that our method to estimate the number of coincident SNPs under simple context effects worked well. In each simulation we obtained 300MB of human

sequence data at random from the reference human genome (GRCh37) and then evolved the sequence under various mutation patterns, in which the mutation rate depended upon the adjacent nucleotides and the type of mutation, to generate simulated human and chimpanzee sequences. Sequences were diverged by 0%, 0.5% or 1% from the reference sequence to generate human and chimp sequences that were 0%, 1% and 2% divergent in total. In the case of the 1% divergence simulations, we used three times as much data to improve the accuracy of our point estimate, since this level of divergence is the most realistic for the divergence between human and chimpanzee (Mikkelsen et al. 2005). Into these sequences we then introduced SNPs according to the same mutation pattern as was used for the divergence step, at the density found in dbSNP – one SNP every 266bp in humans and every 2128bp in chimp. Each SNP was extracted from the simulated human and chimp sequences; we constructed a BLAST database of ~1.1 million human SNPs with 100bp of flanking DNA sequence, and a query dataset of ~140,000 chimpanzee SNPs with 50bp of flanking DNA in each case (with ~3 times as many SNPs for 1% divergence simulations). We ran the BLAST analysis and analysed the output exactly as we had the real data. We performed simulations under four different mutation scenarios. In the first we had no mutation bias for all sites and types of mutation, except for CpG transitions, which mutated at 1, 10, 15, 20 and 30 times the average mutation rate of all other sites. In the second we introduced a simple transition/transversion bias, where transition mutations occurred twice as frequently as transversions at all sites, except for CpG transitions, which again mutated at 1, 10, 15, 20 and 30 times the average mutation rate of all other sites. This pattern of mutation is similar to the broad patterns suggested to occur in the human genome (Zhao and Boerwinkle 2002). In the third model, we incorporated a more complex pattern of mutation as suggested by Zhang *et al* (2007) in which CpG

transversions occur 2.5 times more frequently than non-CpG transversion and non-CpG transitions occur 4.5 times more frequently than non-CpG transversions. Again, CpG transitions occured at 1, 10, 15, 20 and 30 times the average mutation rate of all other sites. Finally, the fourth model incorporates the most complex pattern of mutation rates that are obtained from a study by Hwang and Green (2004). In this model the mutation rate of each site is a function of it's neighbouring nucleotides and the type of mutation, much as is expected to occur in the human genome, and CpG transitions occur ~16.7 times more frequently than the average mutation rate at all other sites.

2.3.5 Paralogous SNPs

In order to rule out the effects of paralogous sequences in generating an excess of human and chimpanzee coincident SNPs we proceeded as follows. We obtained the full fasta sequences for each coincident SNP from the dbSNP website (http://www.ncbi.nlm.nih.gov/projects/SNP/). We then BLASTed these SNPs against the GRCh37 version of the reference human genome, using the parameters outlined in Musumeci *et al* (2010), except we increased the expected value to 0.1 to increase the likelihood that each SNP aligned to the reference sequence. This approach is likely to be more conservative, since it tends to increase the number of genomic hits at coincident SNP sites. Then, following the method in Musumeci *et al* (2010), we extracted only those alignments that contained the SNP position and where at least 20% of the full length SNP sequence had at least 90% sequence identity. Coincident SNP sequences with single hits to the reference genome were considered genuine and therefore not a consequence of different nucleotides being present at the same position in paralogous sequences. For sequences with multiple hits, if only one of the SNP

alleles was present at the SNP position across all alignments, the SNP was considered genuine; if both alleles were present the SNP was considered to be a consequence of substitutions in paralogous sequences and therefore a false positive. Finally, if any of the alleles across matching alignments at the SNP site were not the same as those found in the original SNP, the site was defined as undetermined.

2.3.6 Strand asymmetry

To investigate strand asymmetry, we estimated the mutation rate of the central nucleotide in each triplet by tabulating the number of times each triplet contained a SNP. The direction of mutation was inferred from the frequency; i.e. the minority allele was judged to be the new mutation. We inferred mutation rates across 964 human genes from the Seattle SNPs (SeattleSNPs 2008) and Environmental Genome Projects (NIEHS-SNPs 2008). To investigate which of these genes are expressed in the male germ-line we downloaded gene expression data from the human testis from the study of Ge *et al*. (2005). We obtained raw CEL files of gene expression levels from the NCBI Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/projects/geo/). We normalized the results from the mouse and rat arrays separately using the RMA algorithm (Irizarry et al. 2003) as implemented in Bioconductor (Gentleman et al. 2004). We judged a gene to be expressed within the testis if its expression was above 200 (Su et al. 2004).

2.3.7 Ancestral Polymorphism

In order to model the possible contribution of neutral ancestral polymorphisms to coincident SNPs we proceeded as follows. We started by simulating the coalescent process of a single site in a standard manner by summing random numbers from a series of exponential distributions, in which the mean time to coalescence of m to m-1 lineages is proportional to $1/(m(m-1))$; this sum was multiplied by 4 so the sum would be the time to coalescence in N_e generations. We assumed that the generation time of human and chimpanzee is 25 years (Eyre-Walker and Keightley 1999), that current human and chimpanzee nucleotide diversity levels are 0.001 (Sachidanandam et al. 2001; Venter et al. 2001) and 0.002 (Mikkelsen et al. 2005) respectively, and that the nucleotide diversity of the common ancestor is between 1 and 10 times larger than human diversity levels, consistent with current estimates that the common ancestor of human and chimpanzee may have had an effective population size 4-10 times larger than the current human effective population size (Wall 2003; Burgess and Yang 2008). We also assumed that human and chimpanzee diverged 5 million years ago, which is likely to be conservative (Benton and Donoghue 2007). Since the long-term average effective population size along the lineages leading from human and chimpanzee are unknown, we modeled three scenarios where the average effective population size of human and chimp were 10000, 20000 and 40000. Furthermore, we assumed that the number of chromosomes sampled during the initial discovery of the SNP was 50 and 2 for humans and chimp respectively; the results are very similar if the number of chromosomes sampled varies over a realistic parameter range. We performed each simulation 50,000 times under each set of parameters and the likelihood of an ancestral polymorphism segregating in both species was found by multiplying the proportion of genealogies that

extend back far enough in human and chimpanzee, by the diversity in the common ancestor. This is a slight overestimate of the likelihood that an ancestral polymorphism will be present and sampled in both species, since not all genealogies connecting two human and two chimp lineages will generate a polymorphism that will be inherited by both species. The expected number of ancestral polymorphisms surviving to the present in both species was then compared to half the observed number of coincident SNPs, since approximately one half of coincident SNPs are due to chance alone.

2.3.8 Log-normal model

We estimated the variation in the mutation rate as follows. We start by assuming there is no divergence between humans and chimpanzees so a hypermutable site in humans will also be hypermutable in chimpanzees. Let the average probability of detecting a SNP at a site in humans and chimpanzees be μ_h and μ_c respectively; if μ_h and μ_c are small, the probability at a particular site be $\gamma\mu_h$ and $\gamma\mu_c$, where γ is the relative rate of mutation. Let us assume that γ takes some distribution D(γ) which has a mean of one. The expected number of coincident SNPs is

$$P = \int D(\gamma)\mu_h\mu_c\gamma^2 d\gamma \qquad (2.1)$$

If there is no variation in the mutation rate then this reduces to

$$P_0 = \mu_h\mu_c \qquad (2.2)$$

such that the ratio of the number of coincident SNPs, over the number expected with no variation, is

$$Z = \frac{P}{P_0} = \int D(\gamma)\gamma^2 d\gamma \qquad (2.3)$$

an equation which only depends upon the distribution of γ. We assume that γ is either log-normally distributed, or that it has a two state distribution in which sites can either be hypermutable or normal (see appendix 2.1). We estimate the parameters of the distribution of γ by considering the ratio of the observed number of SNPs over the number expected with simple context effects (i.e. the number expected without cryptic variation in the mutation rate).

This model is unrealistic because we assume that a site does not change its mutation rate; however, hypermutable sites are more likely to change, and this may lead them to become non-hypermutable. Under the log-normal model we assume that once a site changes, its mutation rate is drawn randomly from the log-normal distribution. Let v be the average rate of mutation per unit time in both humans and chimpanzees. Consider a site, in the ancestor of humans and chimpanzees, which currently has a mutation rate $v\gamma$. The probability that the site will remain unchanged along both the human and chimpanzee lineage is

$$Q_u = e^{-2v\gamma t} \qquad (2.4)$$

where t is the time since humans and chimpanzees diverged. The probability that such a site will produce a coincident SNP is

$$P_u = \mu_h \mu_c \int D(\gamma) e^{-2v\gamma t} \gamma^2 d\gamma \qquad (2.5)$$

If the site changes in one of the lineages, then the mutation rates in the two lineages become independent of one another; since the mean of a product is the product of the means, when two random variables are independent, the probability of a coincident SNP at a site which has undergone at least one substitution is

$$P_d = \mu_h \mu_c \int D(\gamma)(1 - e^{-2vt\gamma}) d\gamma \qquad (2.6)$$

The expected number of SNPs with no variation in the mutation rate is still P_0, as given by equation 2.2, so we can write the ratio of the expected number of coincident SNPs with variation over the expected number without variation in the mutation rate as

$$Z = \frac{P_u + P_d}{P_0} = \int D(\gamma) e^{-2v\gamma t} \gamma^2 d\gamma + \int D(\gamma)(1 - e^{-2vt\gamma}) d\gamma \qquad (2.7)$$

This equation depends on the compound parameter $2vt$, which is the average divergence between humans and chimpanzees and the distribution of γ. Since we set the average of the log-normal distribution to one, we need only find the shape parameter of the log-normal distribution.

To estimate the variance associated with simple context effects we calculated the mutation rate of each triplet as above, when correcting simple context effects. We then scaled the mutation rates so the mean across triplets, taking into account their

frequencies in the genome, had a mean of one. We then calculated the variance. This can be compared directly to the variance of the log-normal distribution, which we had also constrained to have a mean of one. We weighted the variance estimates from the CpG and non-CpG sites by the relative frequency of the sites.

2.4 Results

2.4.1 Excess of coincident SNPs

To investigate whether human and chimpanzee SNPs tend to occur at the same sites in the genome, we BLASTed all chimpanzee SNPs against a dataset of human SNPs. This yielded a dataset of 309,158 alignments of 81 base pairs (bp) with the chimpanzee SNP in the central position and a human SNP elsewhere within the alignment. Of these alignments, 11571 have the human and chimpanzee SNP at the same position (figure 2.1); we refer to these as coincident SNPs. This number of coincident SNPs is much greater than the 3817 we would expect if the human SNPs were distributed at random across the alignment, and also much greater than the 6592 we would expect taking into account the influence of the adjacent nucleotides on the mutation rate, henceforth known as "simple" context effects. The observed excess of coincident SNPs is significantly greater than the expected number (ratio of observed over expected with simple context effects = 1.76 with a standard error of 0.02, $p < 0.0001$ under the null hypothesis that the ratio is 1).

a)

		Chimpanzee					
		C/T	G/A	C/A	G/T	C/G	A/T
Human	C/T	3840	11	181	98	197	73
	G/A	14	3708	95	171	189	101
	C/A	226	107	291	3	48	27
	G/T	114	254	0	304	48	16
	C/G	190	194	46	51	217	3
	A/T	81	89	33	19	0	532

b)

		C/T	G/A	C/A	G/T	C/G	A/T
Human	C/T	1.91		1.04	1.19	1.21	0.96
	G/A		1.83	1.24	1.02	1.14	1.40
	C/A	1.23	1.08	4.81		1.28	1.39
	G/T	1.15	1.38		4.95	1.27	0.77
	C/G	1.09	1.14	1.24	1.4	2.79	
	A/T	0.94	1.06	1.79	0.99		15.43

Table 2.1. The pattern of coincident SNPs. a) Number of times a particular SNP in humans is found opposite a particular SNP in chimpanzees. b) The observed number of SNPs over the number expected with simple context effects; for clarity cells in which the observed number of SNPs were less than 20 have been removed because they generate ratios with very large variances. CpG sites are included; see appendix 2.2 for an equivalent table with CpG sites excluded.

This excess is not due to our inability to correct for CpG effects; if we remove CpG dinucleotides from the analysis we observe 5028 coincident SNPs but would only expect 2533 taking into account simple context effects (ratio = 1.98 (0.03); $p < 0.0001$). If we look at the pattern of coincident SNPs, it is evident that almost all the excess is due to the same SNP being present in both humans and chimpanzees, with A-T/A-T SNPs being dramatically over-represented (table 2.1; see appendix 2.2 for the analysis with CpG sites removed).

Although the excess of coincident SNPs is consistent with variation in the mutation rate that is not associated with simple context, there are several other explanations that warrant consideration.

2.4.2 Simulations

In order to check that our method for estimating the expected number of coincident SNPs works well, we performed a number of simulations under various patterns of mutation and different levels of human and chimp divergence, in each case generating simulated SNP datasets that were analysed in the same way as the original analysis. In the basic model all sites were assumed to mutate at the same rate, except for CpG transitions, which mutated at 1, 10, 15, 20 and 30 times the average rate of all other sites (results shown in table 2.2). In the simple transition/transversion model, transition mutations occurred twice as frequently and transversions, except for CpG transitions, which mutated under the same conditions as in the basic model (table 2.3). In the complex transition/transversion model, the patterns of mutation rate mirrored the broad categories as outlined in a study by Zhang *et al* (2007), except for CpG transitions

which again mutated at a rate as outlined in the basic model (table 2.4). Finally, in the Hwang and Green model, sites mutated at a rate as observed in the study by Hwang and Green (2004), which are thought to be the same as those observed in the human genome (table 2.5).

First, the results show that varying the rates of mutation for all mutation types except CpG transitions has little effect on the reliability of the method. All models except the Hwang and Green model show very similar patterns; when all sites are considered there is a tendency to slightly underestimate the expected number of coincident SNPs – this bias increases with increasing CpG transition rate. The bias appears to be maximal when the divergence is 1%. The maximum level of underestimate is such that the observed over expected ratio is 1.14. In contrast, if we consider non-CpG sites the method tends to slightly overestimate the expected number of coincident SNPs. Under the Hwang and Green model, which fixes the rate of CpG transition at ~16.7 times the average rate of mutation at all other sites, results are consistent with other models as the method predicts an observed over expected coincident SNP ratio similar to those observed when the mutation rate for CpG transitions is 15-20 times higher than other sites; there is little or no bias for all sites and a slight overestimate for non-CpG sites. These results tell us that the model of mutation patterns is unlikely to affect the predictions made by our method.

Second, our method works well at all divergences and under all mutation patterns, except when the CpG rate is very high, where the method tends to underestimate the expected number of coincident SNPs. Surprisingly, in the majority of cases the method tends to slightly overestimate the expected number of coincident SNPs when CpG sites

Div	CpG	All sites				Non CpG sites			
(%)	rate	*Obs*	*Exp*	*Ratio*	*95% C.I.*	*Obs*	*Exp*	*Ratio*	*95% C.I.*
0	1x	515	552	0.934	(0.853,1.014)	442	450	0.983	(0.891,1.074)
0	10x	603	613	0.983	(0.905,1.062)	407	408	0.997	(0.900,1.094)
0	15x	685	675	1.015	(0.939,1.091)	359	387	0.926	(0.831,1.022)
0	20x	775	778	0.996	(0.926,1.066)	340	370	0.918	(0.820,1.016)
0	30x	992	880	1.128	(1.058,1.198)	312	316	0.988	(0.878,1.097)
1	1x	1592	1624	0.981	(0.932,1.029)	1282	1290	0.994	(0.940,1.048)
1	10x	1782	1807	0.986	(0.941,1.032)	1163	1162	1.001	(0.944,1.059)
1	15x	2007	1992	1.008	(0.963,1.052)	1083	1105	0.980	(0.922,1.039)
1	20x	2345	2185	1.073	(1.029,1.117)	1073	1057	1.015	(0.955,1.076)
1	30x	2996	2653	1.129	(1.089,1.170)	934	965	0.968	(0.906,1.030)
2	1x	502	507	0.989	(0.903,1.076)	392	410	0.956	(0.861,1.050)
2	10x	511	552	0.926	(0.846,1.007)	310	375	0.826	(0.734,0.918)
2	15x	574	599	0.958	(0.879,1.036)	344	354	0.972	(0.869,1.075)
2	20x	611	593	1.031	(0.949,1.113)	306	313	0.978	(0.869,1.088)
2	30x	730	687	1.062	(0.985,1.139)	268	286	0.937	(0.825,1.049)

Table 2.2: Simulation results for the basic model. Table gives the observed and expected number of coincident SNPs from simulations run with different levels of CpG hypermutability and divergence. The ratio of observed over expected coincident SNPs is also given, together with a 95% confidence interval.

Div	CpG	All sites				Non CpG sites			
(%)	rate	*Obs*	*Exp*	*Ratio*	*95% C.I.*	*Obs*	*Exp*	*Ratio*	*95% C.I.*
0	1x	500	549	0.911	(0.831,0.991)	402	448	0.896	(0.809,0.984)
0	10x	597	603	0.990	(0.910,1.069)	407	417	0.977	(0.882,1.072)
0	15x	636	666	0.955	(0.881,1.029)	372	397	0.938	(0.842,1.033)
0	20x	776	746	1.040	(0.967,1.114)	387	381	1.015	(0.913,1.116)
0	30x	1061	929	1.142	(1.073,1.211)	346	349	0.991	(0.886,1.095)
1	1x	1571	1624	0.967	(0.919,1.015)	1281	1289	0.993	(0.939,1.048)
1	10x	1674	1767	0.948	(0.902,0.993)	1094	1182	0.925	(0.870,0.980)
1	15x	2000	1951	1.025	(0.980,1.070)	1091	1123	0.972	(0.914,1.029)
1	20x	2240	2141	1.046	(1.003,1.090)	1056	1082	0.976	(0.917,1.035)
1	30x	2983	2613	1.142	(1.101,1.183)	1009	980	1.030	(0.966,1.094)
2	1x	504	510	0.987	(0.901,1.074)	409	412	0.993	(0.897,1.090)
2	10x	557	547	1.019	(0.934,1.104)	375	380	0.988	(0.888,1.088)
2	15x	575	593	0.970	(0.890,1.049)	338	364	0.930	(0.830,1.029)
2	20x	702	637	1.102	(1.020,1.183)	351	349	1.006	(0.900,1.111)
2	30x	785	743	1.056	(0.983,1.130)	299	319	0.936	(0.830,1.042)

Table 2.3: Simulation results for the simple transition/transversion model. Table gives the observed and expected number of coincident SNPs from simulations run with different levels of CpG hypermutability and divergence. The ratio of observed over expected coincident SNPs is also given, together with a 95% confidence interval.

Div (%)	CpG rate	All sites				Non CpG sites			
		Obs	*Exp*	*Ratio*	*95% C.I.*	*Obs*	*Exp*	*Ratio*	*95% C.I.*
0	1x	534	557	0.958	(0.877,1.039)	447	446	1.003	(0.910,1.096)
0	10x	612	616	0.994	(0.915,1.072)	426	419	1.018	(0.921,1.114)
0	15x	719	695	1.035	(0.959,1.110)	402	402	0.999	(0.902,1.097)
0	20x	806	776	1.039	(0.967,1.111)	348	379	0.919	(0.822,1.015)
0	30x	1091	973	1.122	(1.055,1.188)	313	349	0.897	(0.798,0.997)
1	1x	1529	1625	0.941	(0.894,0.988)	1205	1272	0.947	(0.894,1.000)
1	10x	1796	1812	0.991	(0.945,1.037)	1168	1184	0.986	(0.930,1.043)
1	15x	2071	1996	1.037	(0.993,1.082)	1114	1132	0.984	(0.926,1.042)
1	20x	2390	2223	1.075	(1.032,1.118)	1078	1080	0.998	(0.939,1.058)
1	30x	3046	2667	1.142	(1.101,1.183)	984	986	0.998	(0.936,1.061)
2	1x	479	513	0.933	(0.849,1.016)	389	410	0.949	(0.855,1.043)
2	10x	552	556	0.993	(0.910,1.076)	375	380	0.987	(0.887,1.087)
2	15x	495	502	0.987	(0.900,1.074)	285	300	0.949	(0.839,1.059)
2	20x	669	654	1.023	(0.946,1.101)	346	350	0.989	(0.885,1.093)
2	30x	906	763	1.187	(1.110,1.265)	296	312	0.948	(0.840,1.056)

Table 2.4: Simulation results for the complex transition/transversion model. Table gives the observed and expected number of coincident SNPs from simulations run with different levels of CpG hypermutability and divergence. The ratio of observed over expected coincident SNPs is also given, together with a 95% confidence interval.

Div (%)	All sites				Non CpG sites			
	Obs	*Exp*	*Ratio*	*95% C.I.*	*Obs*	*Exp*	*Ratio*	*95% C.I.*
0	839	812	1.033	(0.963,1.103)	401	428	0.936	(0.844,1.028)
1	2419	2316	1.045	(1.003,1.086)	1182	1228	0.963	(0.908,1.018)
2	681	685	0.995	(0.920,1.069)	374	400	0.935	(0.840,1.030)

Table 2.5: Simulation results for the Hwang and Green model. Table gives the observed and expected number of coincident SNPs from simulations run with different levels of CpG hypermutability and divergence. The ratio of observed over expected coincident SNPs is also given, together with a 95% confidence interval.

are removed for reasons that are not clear. CpG transitions have been estimated to occur 8-18 times the average rate of all other sites (Hess, Blake, and Blake 1994; Nachman and Crowell 2000; Hwang and Green 2004; Lunter and Hein 2004; Zhang et al. 2007; Elango et al. 2008), meaning that simulations where the CpG transition rate is between 10 and 20 are likely to be the most informative, although the 20 fold increase in CpG transition rate is slightly above the level found in most studies. Furthermore, it has been shown that the level of human and chimp divergence is ~1% (Mikkelsen et al. 2005). Over this CpG transition parameter range and a 1% level of divergence, our method predicts that the observed over expected ratio for all coincident SNPs is between 0.948 and 1.075, and for non-CpG sites the prediction is between 0.925 and 1.015, thus performing well under realistic mutation rates. Perhaps the most realistic set of mutation rates is under the Hwang and Green model, as these rates were estimated using a large dataset and their method incorporates many other parameters such as sequence type and the physical location of substitutions (Hwang and Green 2004). Under this model, at the 1% divergence level our method predicts that the observed over expected ratio for all coincident SNPs is 1.045, and for non-CpG sites the prediction is 0.963. As a consequence, it appears that our method introduces a very slight bias when predicting the number of coincident SNPs, however in the case of non-CpG coincident SNPs the method appears to be conservative under the most realistic conditions.

2.4.3 Errors in dbSNP

In order to identify the number of human and chimpanzee SNPs that are coincident we used single nucleotide polymorphism data obtained from the online database dbSNP. The data quality of this resource has been questioned (Mitchell et al. 2004; Musumeci et

al. 2010) and under certain conditions it may be possible that the excess of coincident SNPs is a consequence of these errors. First, it has been suggested that up to ~15-17% of the entries in dbSNP are false positives, occurring due to sequencing errors arising from base calling software, particularly at heterozygous sites (Mitchell et al. 2004). Should errors that arise due to base calling be more prevalent for coincident SNPs, we might expect the frequency of coincident SNPs that are confirmed only in heterozygous individuals to be higher than for other SNPs. In order to test this we obtained genotype information from dbSNP for coincident SNPs and compared the proportion of cases where the minor allele of the SNP was only confirmed in heterozygous individuals with the same number of random SNPs from the same resource. We were able to obtain genotype information for 7025 coincident SNPs; amongst these 0.355 were confirmed only in heterozygous individuals (95% confidence interval: (0.341,0.369), assuming the number of SNPs only confirmed in heterozygotes is Poisson distributed) and for the same number of random SNPs, this value is 0.398 (0.384,0.413). Since the proportion of SNPs confirmed only in heterozygotes is significantly higher in random SNPs than coincident SNPs ($p<0.05$), there is no evidence to suggest that coincident SNPs are more prone to base calling errors than other SNPs.

Second, it has been suggested that a proportion of SNPs in dbSNP may be false positives due to substitutions occurring within paralogous regions of the human genome (Fredman et al. 2004; Musumeci et al. 2010), and that this may account for ~8.32% of entries in the database (Musumeci et al. 2010). False positive SNPs may occur in this fashion if researchers are unable to discriminate between two paralogous regions of the genome; substitutions in these regons would then be observed as a SNP during sequencing. Paralogous sequences may be problem in our analysis if a substitution

occurred in a duplicated region prior to the split of human and chimpanzee, as a false positive SNP would then be called at the same site in both species. In order to consider whether paralogous sequences contribute to the excess of coincident SNPs, we repeated the analysis of Musumeci *et al* (2010) and BLASTed all coincident SNP sequences against the human reference genome and examined all cases where there were multiple matches. Single hit sequences were regarded as genuine SNPs. If sequences with multiple matches contained both alleles called at the SNP site we categorized the coincident SNP as a potential false positive. Alternatively, if an allele was present in one or more alignment that was not the same as one of the alleles in the original SNP, the SNP was categorized as undetermined.

In total, out of the 11,571 coincident SNPs found between human and chimpanzee, 9,611 had a single match to the reference genome, 233 had multiple matches but only a single allele was found at the SNP site across all alignments, 269 had multiple matches to the reference genome and both SNP alleles were found at the SNP site and 95 coincident SNPs were undetermined. Furthermore, we were unable to match 1,363 coincident SNP sequences to the reference genome due to the parameters of the BLAST search; these sequences contained some low complexity or interspersed repeats that interrupted the BLAST extension. Out of the coincident SNPs that had at least one match to the reference genome, ~3.6% are potential false positives due to variant alleles in paralogous sequences, which is lower than the ~8.32% found in the study by Musumeci *et al* (2010). However, it may be the case that a high proportion of the coincident SNPs that we were unable to match to the reference genome are a consequence of substitutions in paralogous sequences. It is worth noting that the vast majority of these SNPs have been mapped to single locations on the dbSNP website, by

using a less stringent BLAST criteria that allows seeding of shorter length sequences that then extent into repeat regions (Kitts and Sherry 2010). Under a conservative approach, if we assume that all coincident SNPs that we have not matched to the reference genome are in fact a consequence of substitutions in paralogous sequences, and that false positives occur at non-coincident SNP sites at the same rate as found in the Musumeci *et al* analysis (2010), we estimate that the observed over expected ratio of coincident SNPs reduces to 1.63 (standard error 0.016) from 1.76. Under the same criteria, the observed over expected ratio for non-CpG coincident SNPs reduces to 1.81 (standard error 0.028) from 1.98.

It may also be possible that some coincident SNPs are a consequence of substitutions occurring in paralogous regions that have not been identified in the reference genome, for example, in copy number variants. It is not possible to determine the extent of this phenomenon using the method above, however if a high proportion of coincident SNPs are a consequence of substitutions in paralogous regions, we might expect the average minor allele frequency at coincident SNP sites in humans to be higher than other SNP sites since, for example, variant alleles picked up at two different positions in the genome would likely be at a minor allele frequency of ~0.5. To investigate this possibility we obtained the minor allele frequencies at coincident SNP sites from dbSNP and compared these to the minor allele frequencies at the same number of randomly chosen SNPs from the same resource. In total we were able to obtain allele frequency information for 7801 of the coincident SNPs, which have an average minor allele frequency of 0.274 (95% confidence interval: (0.270, 0.277)); the same number of randomly chosen SNPs have an average minor allele frequency of 0.271 (0.267,0.274), which is not significantly different to that observed at coincident SNPs (t-test, p=0.241).

As the average minor allele frequency of coincident SNPs is only slightly higher than that of the same number of randomly chosen SNPs, it seems unlikely that a high proportion of coincident SNPs are a consequence of undetected paralogous regions, or indeed detected paralogous regions, in the human genome.

2.4.4 Strand asymmetry

In correcting for simple context effects, we have also made two assumptions; we have assumed that the pattern of mutation is the same on the two strands of the DNA duplex, and we have assumed that context effects are the same across the genome. As a consequence of these assumptions, we could be underestimating the expected number of coincident SNPs. For example, let us imagine that the central nucleotide in the triplet AAA has a high mutation rate on one strand, say the transcribed strand, and a low mutation rate on the other strand, but that the pattern is the opposite for the triplet CCC (note that when we refer to the mutation of a triplet, we are referring to the mutation rate of the central nucleotide). Because the relative mutation rates of AAA and CCC depend upon which strand we are considering, we would tend to underestimate the expected number of coincident SNPs.

The pattern of mutation is known to differ between the two DNA strands in a manner that depends upon transcription (Green et al. 2003; Gibbs et al. 2004). However, what is important for our analysis is whether the relative mutation rates of the triplets differs between strands; it is the relative, rather than the absolute rate that matters, because for each alignment we calculate the chance of a coincident SNP relative to the chance that the human SNP occurs at one of the other triplets in the sequence. To investigate this,

we estimated the mutation rate of the central nucleotide in each triplet for a set of human genes for which we knew the direction of transcription; we also considered a subset of these genes known to be expressed in the testis.

In agreement with Green *et al*. (2003), we observe a 25% excess of A->G transitions over T->C transitions; however, we did not observe an excess of G->A transitions over C->T transitions, even in our testes expressed genes. Crucially for our analysis, the mutation rate of each triplet is highly correlated to its reverse-complement triplet for all genes (Pearson correlation coefficient r = 1.00 for all triplets, r = 0.85 without triplets containing CpGs; appendix 2.6a) and for genes expressed in the testes (r = 0.99 for all triplets, r = 0.75 without triplets containing CpGs; appendix 2.6b); genes expressed in the testes are expressed in the male germ-line, where any strand asymmetry in the pattern of mutation will have an evolutionary effect. It therefore seems unlikely that strand asymmetry in the pattern of mutation is leading to an underestimate of the expected number of coincident SNPs.

2.4.5 Patterns of mutation

The excess of coincident SNPs could also be due to variation in the pattern of mutation across the genome for reasons similar to those given for strand asymmetry; if the relative rate at which each triplet mutates differs between genomic regions, we will underestimate the expected number of coincident SNPs. Since such variation in the pattern of mutation might be expected to generate differences in base composition, we divided our dataset of alignments according to their GC content and estimated the mutation rate of the central nucleotide in each triplet in the chimpanzee sequence using

the human sequence to infer the ancestral sequence. The relative rates of mutation inferred from the sequences in the upper and low GC content quartiles are highly correlated to each other (r = 0.99 using all triplets; r = 0.88 excluding triplets involving CpGs; appendix 2.7), which suggests that triplets that are highly mutable in high GC content sequences also tend to be highly mutable in the low GC content sequences. It therefore seems unlikely that we are underestimating the expected number of coincident SNPs because of variation in the pattern of mutation. As expected, we find a significant excess of coincident SNPs in both the upper and lower GC quartile datasets, although the excess of coincident SNPs appears to be slightly stronger in GC-poor DNA (appendix 2.3).

2.4.6 Ancestral polymorphism

The excess of coincident SNPs could potentially be due to inheritance, in humans and chimpanzees, of polymorphisms that were present in their last common ancestor. In order to attempt to quantify the contribution of ancestral polymorphisms to coincident SNPs we performed a number of simulations in which we modeled the genealogy of a site in both human and chimp and then calculated the likelihood that these genealogies extend back as far as the estimated divergence of human and chimp 5 million years ago; we assume these polymorphisms are neutral, since it is not possible to quantify the contribution of balanced polymorphisms. The likelihood is then compared to the random expectation for the number of coincident SNPs present in current populations, since the excess of coincident SNPs is greater than two fold without incorporating context effects, to give an estimation of the contribution of ancestral polymorphism to coincident SNPs. We make a number of assumptions about the human and chimp

populations and their demographic history – we assume an average generation time of 25 years (Eyre-Walker and Keightley 1999), that the current human and chimpanzee diversity is 0.001 (Sachidanandam et al. 2001; Venter et al. 2001) and 0.002 (Mikkelsen et al. 2005) respectively, and that the number of chromosomes sampled during the initial discovery of the SNP is 50 and 2 for human and chimp respectively; the results are very similar if the number of chromosomes sampled varies over a realistic parameter range. Furthermore, we assume that the diversity in the common ancestor of human and chimpanzee is 1, 5 and 10 times the current human diversity; this is in line with expectations that the effective population size of the common ancestor is between 4 and 10 times that of the current human effective population size (Wall 2003; Burgess and Yang 2008). The likelihood of a common polymorphism being present in both human and chimpanzee also depends on the long-term average effect population size along the lineages leading from the common ancestor to human and chimp. The current effective population sizes for human and chimp are thought to be around 10,000 (Yu et al. 2001) and 20,000 (Hey 2010) respectively, however we have very little information about the long-term effective population sizes of these two species, and values could vary dramatically depending on whether the decrease from the larger effective population size assumed for the common ancestor was linear, exponential or even involved a population bottleneck. Consequently, we modeled average long-term effective population sizes of 10,000, 20,000 and 40,000 for both human and chimpanzee.

The results of the simulations are presented in table 2.6. The contribution of ancestral polymorphism to coincident SNPs varies dramatically, depending on assumptions about the long-term average effective population size between human, chimp and their common ancestor, and the nucleotide diversity present in the common ancestor. When

Common Ancestor Diversity	Human Ne	Chimp Ne	Contribution to coincident SNPs (%)
1x	10000	10000	0.0001
1x	10000	20000	0.0159
1x	10000	40000	0.1844
1x	20000	10000	0.0083
1x	20000	20000	1.4902
1x	20000	40000	17.247
1x	40000	10000	0.1035
1x	40000	20000	18.472
1x	40000	40000	213.79
5x	10000	10000	0.0004
5x	10000	20000	0.0797
5x	10000	40000	0.9223
5x	20000	10000	0.0417
5x	20000	20000	7.4508
5x	20000	40000	86.237
5x	40000	10000	0.5174
5x	40000	20000	92.358
5x	40000	40000	1069.0
10x	10000	10000	0.0009
10x	10000	20000	0.1594
10x	10000	40000	1.8446
10x	20000	10000	0.0835
10x	20000	20000	14.902
10x	20000	40000	172.47
10x	40000	10000	1.0348
10x	40000	20000	184.72
10x	40000	40000	2137.9

Table 2.6: The contribution of ancestral polymorphism to the excess of coincident SNPs under different long-term average effective population sizes for human and chimp and varying nucleotide diversity in the common ancestor.

the long-term average effective population sizes are assumed to be 10,000 in both human and chimp, the contribution of ancestral polymorphisms to the excess of coincident SNPs is insignificant, with estimates well below 0.001% in all cases. If the average effective population sizes are assumed to be 20,000 for both human and chimp, the contribution of ancestral polymorphisms to the excess of coincident SNPs varies between ~1.5% and ~14.9%, thus making a reasonable contribution. However, if we assume that the average effective population sizes for both species is 40,000, ancestral polymorphisms can explain between ~214% and ~2138% of the excess of coincident SNPs. This scenario is clearly unlikely to be realistic, because we would expect to observe far more coincident SNPs than we actually do. The estimates reported under all models are likely to represent maximum values since we have been conservative in our estimates of the divergence time between human and chimpanzee (Benton and Donoghue 2007), and we have not taken into account the fact that many genealogies prior to the species split will not be such that they yield a polymorphism in each species, or that the mutation has to occur in a specific part of the genealogy to generate two polymorphisms.

Our simulations show that ancestral polymorphisms could make a substantial or inconsequential contribution to coincident SNPs, depending on the model of demography used in simulations and ultimately the long-term average effective population size along the lineages leading to human and chimp from the common ancestor. Unfortunately, as we have little information about the average effective population sizes we are unable to rule a contribution to the excess of coincident SNPs from ancestral polymorphisms from a theoretical perspective. However, it seems unlikely that ancestral polymorphism can explain much of the excess of coincident

SNPs as two additional lines of evidence suggest that this is not the case. First, we repeated the analysis using human and macaque SNPs. Since these two species diverged more than 23-34 million years ago (MYA) (Benton and Donoghue 2007), as opposed to the 6-10 MY that separates human and chimp (Benton and Donoghue 2007), one would expect very few polymorphisms to be shared between human and macaque. However, in this dataset we also see a significant excess of coincident SNPs whether we consider all sites (ratio = 1.64 (0.19) $p < 0.001$), or non-CpG sites (1.51 (0.26) and $p < 0.05$). The excess is not due to problems with our method; we repeated simulations as above, using the Hwang and Green model (2004) and setting the level of divergence to 6%, which is roughly the level of divergence estimated between human and macaque (Gibbs et al. 2007). The observed over expected ratio for all sites is 1.001 (95% confidence interval (0.922, 1.080)), which is a slight under-estimate of the number of coincident SNPs expected under simple context effects, however the observed over expected ratio for non-CpG sites is 0.929 (0.833,1.024), which is an over-estimate of the number of coincident SNPs expected under simple context effects. Second, the pattern of coincident SNPs (table 2.1) is inconsistent with ancestral polymorphism. All four of the possible transversion SNPs are approximately equally common amongst SNPs in general (proportion of transversions amongst human SNPs: G/T = 0.092, C/A = 0.091, C/G = 0.088, A/T = 0.075; transitions: C/T = 0.33, G/A = 0.33). We would therefore expect a G-C SNP in chimps to be coincident with a G-C SNP in humans approximately equally often as an A-T SNP in humans is coincident with an A-T SNP in chimps. However, we see distinct biases, with coincident A-T/A-T SNPs being much more common than the other transversions and a bias towards transversions in general. Although we cannot explain this pattern of mutation, it is clearly inconsistent with ancestral polymorphisms at neutral sites.

2.4.7 Natural Selection

It is also possible for the apparent excess of coincident SNPs to be due to selection; if some regions of the genome are under selection, then we expect them to have a low density of SNPs, because many SNPs will be removed as they are deleterious. As a consequence, SNPs will be clustered between these regions, causing an apparent excess of coincident SNPs. This seems an unlikely explanation, since the vast majority of our data is intergenic and intronic (98% and 99% of the human and chimpanzee SNPs in our BLAST databases, respectively), and although selection is known to act within these regions, it is thought to only affect a small percentage of sites (Waterston et al. 2002; Dermitzakis, Reymond, and Antonarakis 2005; Asthana et al. 2007). Furthermore, if selection was causing an excess of coincident SNPs, we would expect SNPs to be clustered generally, but this is not observed (figure 2.1). There is a small excess of human SNPs adjacent to the chimpanzee SNP, but this is a consequence of CpG effects – the chimpanzee SNP is disproportionately likely to occur within a CpG, which means that a human SNP is also likely to occur at the same site, or at an adjacent site. If we remove CpGs, this slight excess of adjacent SNPs disappears (appendix 2.5). Otherwise there is no tendency for SNPs to cluster.

2.4.8 Other context effects

It therefore seems that the majority of the excess of coincident SNPs is a consequence of variation in the mutation rate that is not associated with simple context effects, variation in context effects between strands or regions of the genome, or natural selection. The question therefore arises as to whether the variation in the mutation rate

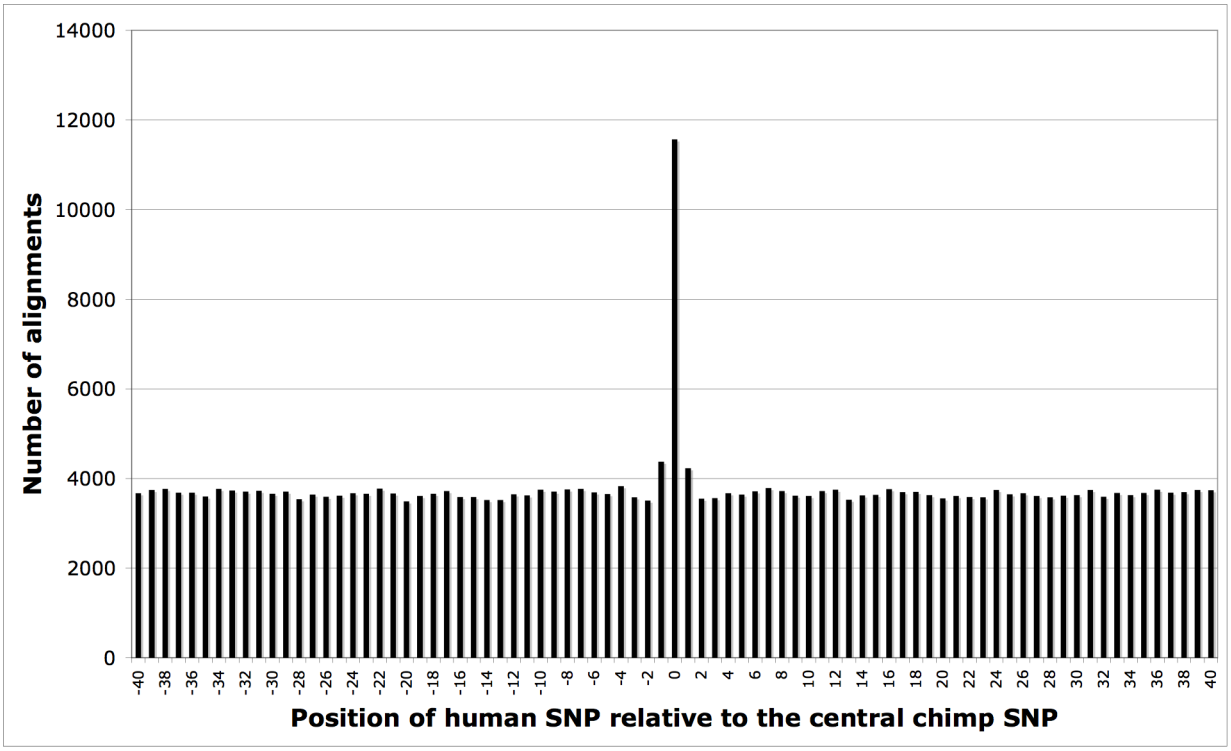
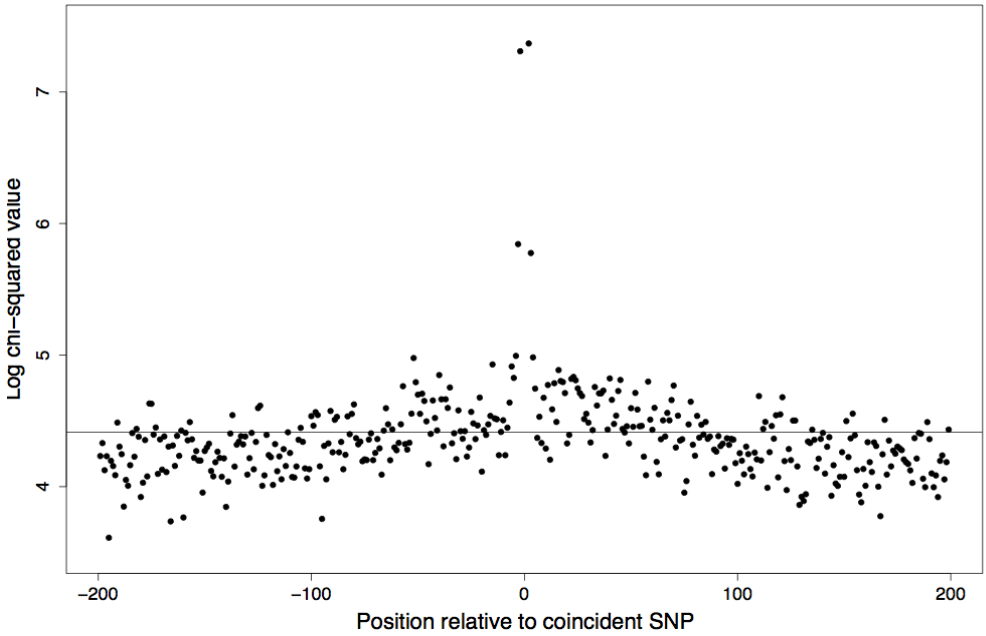


Figure 2.1. The number of human SNPs at each site of the human-chimpanzee alignments used in the analysis.

is associated with other contexts that are distant from the target site, degenerate in nature, or sufficiently complex to be difficult to discern. It should be noted that simple context effects beyond the adjacent nucleotides (e.g. 1bp removed from the target site) are not responsible for the excess. Although these effects exist (Zhao et al. 2003), they are much smaller than the effects of the adjacent nucleotides, which themselves have a relatively modest effect if we remove CpGs; e.g., the expected number of non-CpG coincident SNPs is 2115 if we ignore adjacent nucleotide effects, and it is 2533 if we include these effects.

To investigate whether there are other, more complex context effects, we tabulated the frequency of each triplet at each site in the alignments containing coincident SNPs, and a similar size dataset of alignments with non-coincident SNPs. Surprisingly, we found significant heterogeneity in triplet frequencies that extends to about 80bp either side of the coincident SNP (figure 2.2a); i.e. the relative frequencies of the triplets at sites close to the coincident SNP are different to the average across the alignments. In contrast, if we consider alignments without a coincident SNP, but with a chimpanzee SNP, we only see significant heterogeneity in triplet frequencies within 10bp either side of the SNP (figure 2.2b). Despite the heterogeneity in triplet frequencies surrounding a coincident SNP we could discern very few patterns in the triplets that are over- or under-represented. The only conspicuous pattern is an excess of TTT triplets upstream and AAA triplets downstream of coincident SNPs. However, this seems to explain little of the overall excess of coincident SNPs. If we repeat the analysis but remove all cases in which there is a run of three or more nucleotides, of any type, with or without SNPs within them, from our alignments, we find 8536 alignments with a coincident SNP versus an expected number of 4434, taking into account simple context effects

a)



b)

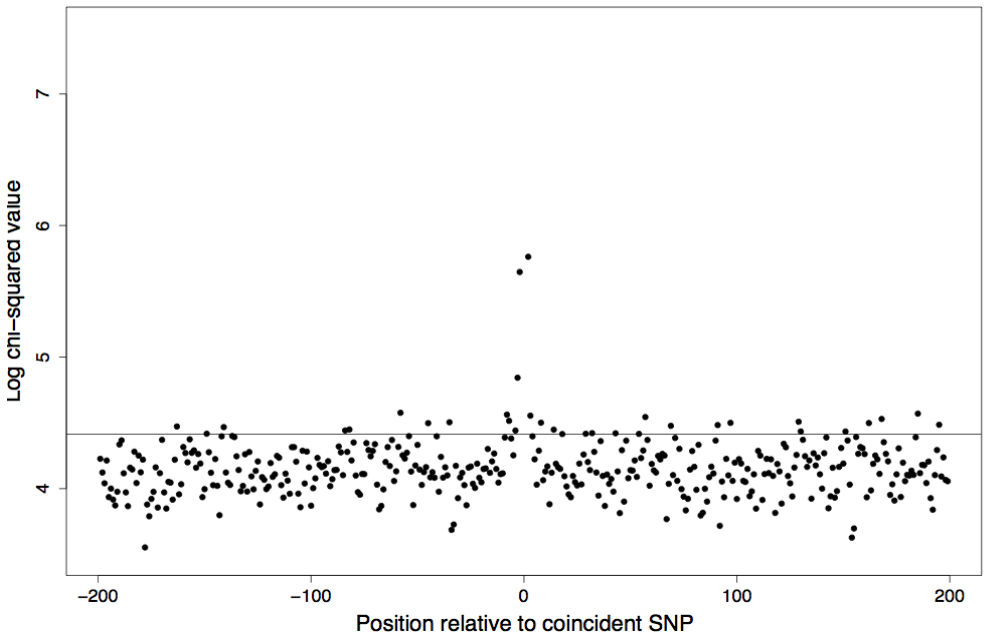


Figure 2.2. Heterogeneity in triplet frequencies. Figure gives the log value from a chi-square test of heterogeneity of triplet frequencies at each site of the human-chimpanzee alignment versus the average triplet frequencies across the whole alignment for (a) alignments containing a coincident SNP, and (b) alignments without a coincident SNP, but with a chimpanzee SNP at the central position. The line marks the point above which 5% of the chi-square values are expected to fall by chance alone. The chi-square values are not given for the central three sites because the presence of the chimpanzee SNP in the centre of the alignment means that triplets cannot be counted at positions 0, 1 and -1.

(ratio = 1.93 (0.02) $p<0.0001$). Considering pentamers, rather than triplets, also fails to reveal any context that is associated with coincident SNPs, except for the α-polymerase pause site motif, TG(A/G)(A/G)(G/T)(A/C), which has been suggested as a hypermutable motif (Todorova and Danieli 1997; Templeton et al. 2000). However, we only observe an excess of α-polymerase pause sites immediately downstream of coincident SNPs, and the total number of coincident SNPs explained by this motif is trivial (2.2%).

2.4.9 Quantification

To quantify the level of cryptic variation in the mutation rate, we fit two models to the ratio of the observed number of coincident SNPs over the number expected with simple context effects. In the first model, we assumed that the variation in the mutation rate was log-normally distributed; in the second, we assumed that there were two types of sites - normal and hypermutable sites. These models give qualitatively similar estimates of the variation, so we only discuss the log-normal model in detail, because this is a model with a single parameter (details of the two-rate model are given in appendix 2.1 and results are in appendix 2.4). Since our method for controlling for simple context effects tends to underestimate the expected number of coincident SNPs when we have CpG sites, we concentrate on non-CpG sites. Also, since we are unsure about the contribution of ancestral polymorphisms and substitutions in paralogous sequences to the excess of coincident SNPs, we use the original estimate of the observed over expected ratio of coincident SNPs (1.98) to represent the maximum range of the effect of cryptic variation in the mutation rate. We fit two sub-models to our data. In the first, we assume that the mutation rate of a site is invariant in both humans and chimpanzees.

Under this "static" model we estimate the shape parameter of the log-normal to be 0.83 (95% CIs of 0.81, 0.84) for non-CpG sites. However, this model may not be realistic, since we might expect sites with high mutation rates to destroy themselves; e.g. if a site has a high rate of C->T mutation, then it will rapidly become fixed for T and therefore become non-hypermutable. We therefore also fit a model in which the time a site remains at a certain mutation rate depends upon that mutation rate, assuming an average divergence between humans and chimpanzees of 0.92% for non-CpG sites (Mikkelsen et al. 2005). Under this model we estimate slightly higher levels of cryptic variation: we estimated the shape parameter to be 0.85 (0.83, 0.87) – higher shape parameters mean more variation. The level of variation that these distributions represent is considerable; with a shape parameter of 0.85 the fastest 5% of sites mutate at least 16.4-fold faster than the slowest 5% of sites. This level of variation in the mutation rate is greater than the variation associated with simple context: the variance due to simple context, including CpGs, is 0.59, whereas the variance due to cryptic variation at non-CpG sites is 1.05. However, this large difference in variance might be due to the model. If we consider a simple two-state model in which sites are either hypermutable or normal, and constrain the proportion of hypermutable sites to be 2%, the proportion of sites that are involved in CpGs in the human genome (Lander et al. 2001), then we estimate that hypermutable sites would have to mutate 9.3-fold faster than normal sites to explain the excess of coincident SNPs. This is similar to 10-20 fold higher rate that CpGs mutate (Hwang and Green 2004; Mikkelsen et al. 2005).

2.5 Discussion

We have shown that there is an excess of sites that have a SNP in both the human and chimpanzee genomes. We demonstrated that this is not due to natural selection, local context effects or problems with our method; although our method tends to slightly under-estimate the expected number of coincident SNPs across all sites, which may contribute to part of the excess, it is conservative under most realistic conditions when CpG sites are excluded, where we still observe a massive excess of coincident SNPs. It also seems unlikely that paralogous sequences have contributed significantly to the excess of coincident SNPs, since relatively few of our coincident SNPs map to multiple locations of the genome, and the average minor allele frequency at coincident SNPs is no higher than at other SNP sites. We have shown that it is theoretically possible that ancestral polymorphisms could contribute substantially to the level of coincident SNPs, however without a greater knowledge of the demographic history of human and chimpanzee and ultimately the long-term average effective population size of the two species, it is difficult to make any firm conclusions about these simulations. Furthermore, two additional lines of evidence suggest that ancestral polymorphism does not contribute substantially to the excess of coincident SNPs; we observe an excess of coincident SNPs between two more distantly related species, where shared polymorphisms are likely to be extremely rare, and the pattern of mutation is inconsistent with ancestral polymorphism at neutral sites contributing all of the excess; we see large biases in the proportions of different types of transversions amongst coincident SNPs. Given the caveats above, it therefore seems likely that the majority of the excess of human and chimpanzee coincident SNPs is due to variation in the mutation rate that is not associated with simple context effects and is cryptic in nature.

We also show that triplet frequencies surrounding sites with coincident SNPs are highly non-random, but we have been unable to discern any specific motifs in these regions. This suggests that there are probably complex context effects that extend some distance from the site they affect. Furthermore, we show that there has to be considerable variation in the mutation rate to explain the observed excess of coincident SNPs.

The presence of such cryptic variation in the mutation rate is perhaps not surprising given the evidence that some sites in the human mitochondrial genome are hypermutable. Hypermutation had long been suspected based on the excess of homoplasies in human mtDNA phylogenies (e.g., see (Meyer, Weiss, and von Haeseler 1999)) and although such an excess could be due to hypermutation or recombination (Eyre-Walker, Smith, and Maynard Smith 1999), two recent analyses have provided convincing evidence that the excess is due to hypermutation. Stoneking (2000) showed that mitochondrial mutations in human pedigrees tend to occur at sites that have high levels of homoplasy, and Galter *et al*. (2006) have recently shown that synonymous mitochondrial SNPs tend to occur at the same positions in different species.

However, many of the hotspots in mtDNA appear to be due to strand slippage type mutational mechanisms (Kunkel and Soni 1988; Malyarchuk and Rogozin 2004), this does not appear to be case for the cryptic variation in the mutation rate in nuclear DNA that we describe here. There are two slippage mechanisms that can operate: template strand and primer strand dislocation. Template strand dislocation is controlled for in our simple context analysis, and primer strand dislocation is controlled for in the analysis of homonucleotide runs.

It has also been shown recently that the mutation rate is elevated close to insertion and deletion mutations in the nuclear genomes of several eukaryotes, including humans (Tian et al. 2008). However, it seems unlikely that this process is generating the excess of coincident SNPs. Indels appear to increase the rate of mutation, but not at specific sites; rather the mutation is elevated close to an indel and this elevation in the mutation rate declines over several hundred nucleotides. This would menifest itself as general tendency for SNPs to cluster, which we do not observe (figure 2.1, appendix 2.5); we only observe an excess of coincident SNPs and a small excess of adjacent SNPs. Furthermore, humans and chimpanzees would both have to have segregating indels in the same locality to generate an excess of coincident SNPs.

Although we have observed an excess of coincident SNPs, we do not know the specific mechanism that generates the excess. Consequently, it is interesting to consider whether there are SNPs in the human genome that are coincident with SNPs in both the chimpanzee and Macaque genomes, as this may shed some light on the time frame over which such a mechanism would need to operate, the locations of sites that are particularly prone to cryptic hypermutation and ultimately the origin of such events. However, due to the limited amount of data available for the Macaque genome, we are unable to find any occurrences of SNPs that are coincident across all three species. This is perhaps not surprising, since we only observe a small number of coincident SNPs between human and macaque, sites would need to be ~300 more mutable than the average site in order to generate just one co-occurrence between human, chimpanzee and macaque, assuming a random distribution of such events.

Over the last few years, DNA sequence analysis has revealed that the mutation process is highly complex, varying between different parts of the genome and between different sites. Unfortunately we do not yet understand many of these patterns.

Acknowledgements. We are very grateful to Vini Pereira for help with the gene expression analysis, and to Nina Stoletzki, Peter Keightley and two referees for comments. AH and AEW were funded by the BBSRC, EL and AEW by the European Community and AEW by the National Evolutionary Synthesis Center.

3. The Genomic Distribution and Local Context of Coincident SNPs in Human and Chimpanzee

3.1 Abstract

We have previously shown that there is an excess of sites that are polymorphic at orthologous positions in humans and chimpanzees and that it seems likely that the majority of this excess is due to cryptic variation in the mutation rate. We showed that this might be a consequence of complex context effects since we found significant heterogeneity in triplet frequencies around coincident SNP sites. Here we show that the heterogeneity in triplet frequencies is not specifically associated with coincident SNPs, but is instead driven by base composition bias around CpG dinucleotides. As a result, we suggest that cryptic variation in the mutation rate is truly cryptic, in the sense that the mutation rate does not appear to depend on any specific primary sequence context. Furthermore, we propose that the patterns around CpG dinucleotides are driven by the mutability of CpG dinucleotides in different DNA contexts. We also show that the genomic distribution of coincident SNPs is non-uniform and that there are some subtle differences between the distributions of single and coincident SNPs. Furthermore, we identify regions that contain high numbers of coincident SNPs and suggest that one in particular, a region containing the gene PRIM2, may be under balancing selection.

3.2 Introduction

There is variation in the mutation rate over a number of different scales in the human genome; on a local scale there are hypermutable sites (Blake, Hess, and Nicholsontuell 1992; Zhao et al. 2003; Hwang and Green 2004), and more broadly, large genomic regions and whole chromosomes can vary in their mutation rate (Matassi, Sharp, and Gautier 1999; Williams and Hurst 2000; Lercher, Williams, and Hurst 2001; Li, Yi, and Makova 2002; Gaffney and Keightley 2005). What makes a region or site have a higher or lower mutation rate is poorly understood, except in the case of CpGs where cytosine can become methylated and unstable, leading to a higher rate of mutation (Coulondre et al. 1978; Bird 1980). However, understanding the factors that dictate the mutation rate is important since they influence human disease and our understanding of evolution.

In the previous chapter we showed that there is an excess of coincident SNPs, sites that have a SNP in both humans and chimpanzees, and that this excess could not be explained by the known influence of adjacent nucleotides on the mutation rate. We also showed that the excess is not a result of selection; positive selection tends to remove variation from the population through rapid fixation of beneficial alleles, and negative selection, in which removal of variation may result in a general clustering of single SNPs in non-coding regions, was not observed. Furthermore, the excess of coincident SNPs is not a consequence of us mis-inferring mutation rates in different parts of the genome; we show that mutation rates correlate in GC-rich and GC-poor regions of the genome, and we also observe a significant excess of coincident SNPs in both sequence contexts. We were unable to rule out a contribution from ancestral polymorphism or substitutions in paralogous sequences that have been misinferred as SNPs, however it

seems unlikely that these two factors contribute substantially to the excess of coincident SNPs; we observe a significant excess of coincident SNPs in comparisons with more distantly related species (human and macaque), and it is unlikely that SNPs would be preserved over this time frame. Furthermore, the pattern of mutation is inconsistent with ancestral polymorphism as we observed a bias in the proportions of different types of transversions at coincident SNP sites. Similarly, we showed that only a small percentage of coincident SNP sequences had both SNP alleles present amongst multiple matches to the human reference genome and even under a conservative approach of removing all coincident SNP sequences that are potential false positives and those that do not match to the reference sequence, the observed over expected ratio for coincident SNPs only reduces to 1.63 from 1.76 for all coincident SNPs and to 1.81 from 1.98 for non-CpG coincident SNPs. Furthermore, the average minor allele frequency at coincident SNP sites is not different to other SNP sites; should a significant number of coinicident SNPs be driven by substitutions in paralogous regions, we might expect the average minor allele frequency to be higher for coincident SNPs. We therefore proposed that there is cryptic variation in the mutation rate. However, despite the evidence that this variation in the mutation rate is not due to simple context effects (i.e. the adjacent nucleotides), we did show that triplet frequencies are significantly heterogeneous to approximately 80bp either side of the coincident SNP.

Here, we investigate both the local and genomic context of human and chimp coincident SNPs. Although many of the coincident SNPs we have identified will not be a consequence of cryptic variation in the mutation rate, as approximately half are due to chance alone and others may be a consequence of ancestral polymorphism or substitutions in paralogous sequences, we consider the entire dataset as we have no way

of differentiating coincident SNPs which are caused by chance alone, and those caused by some other process. We now find that the heterogeneity in triplet frequencies is not specifically associated with coincident SNPs, but is instead associated with patterns of base composition around CpG dinucleotides. As a result, we suggest that cryptic variation in the mutation rate is complex, in the sense that the mutation rate does not appear to depend on any specific context. We also show that the genomic locations of coincident SNPs are non-uniform, and that there are subtle differences in the distributions of single SNPs compared to coincident SNPs across the genome.

3.3 Materials and Methods

3.3.1 Coincident SNPs

In our original analysis (chapter 2) we investigated whether human and chimpanzee SNPs tended to occur at the same site in the genome by BLASTing all chimpanzee SNPs found in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) against a dataset of human SNPs from the same resource. We obtained >300,000 81bp alignments that contained both a human and a chimpanzee SNP, and in 11,571 cases the human and chimpanzee SNPs occurred at orthologous positions. We showed that this number was significantly more than we would expect to occur by chance if SNPs are randomly distributed along our alignments, even after taking into account that certain sites are more likely to contain a SNP due to the effects of neighbouring nucleotides on the mutation rate.

3.3.2 Sequence contexts

In order to investigate heterogeneity in triplet frequencies around coincident SNPs and dincucleotides we analysed the 200bp either side of each of the coincident SNPs identified in our previous analysis (chapter 2), except for a few changes due to the random selection of BLAST alignments, together with an equal number of randomly chosen instances of each dinucleotide. Random dinucleotides were obtained from the entire human genome sequence from the Ensembl database (http://www.ensembl.org/index.html - build 55). We split our dataset of coincident SNPs into two groups, CpG and non-CpG coincident SNPs. A SNP was designated as CpG if the site, or any of the alleles at the site, would yield a CpG dinucleotide.

To investigate whether triplet frequencies are significantly heterogeneous around coincident SNPs and dinucleotides we proceeded as follows. We tabulated the frequency of each triplet at each site relative to the coincident SNP or dinucleotide – e.g. to investigate heterogeneity in triplets 10bp upstream of coincident SNPs, we tabulated the frequency of triplets where the central nucleotide is 10bp upstream of a coincident SNP across all our 11,571 sequences containing a coincident SNP. We then summed the number of each triplet across all sites and divided this by the total number of sites to yield the average expected number of triplets at all sites. Whether the observed values were significantly different to the expected values was assessed using a standard chi-square test. To investigate whether the heterogeneity in triplet frequencies could be explained in terms of trends in base composition we calculated the expected frequency of each triplet from the average nucleotide composition at each site; e.g. to calculate the expected frequency of CTG at position +10, we would multiply the

frequency of C at position +9 by the frequency of T at position +10 and the frequency of G at position +11.

To analyse the sequence context around each type of dinucleotide we obtained 2000 instances of each type of dinucleotide, flanked by 5000bp either side, at random from the human genome. Any sequences that contained possible CpG islands were removed; CpG islands were identified by following the method as outlined by Takai and Jones (2002). For each dataset we lined up the sequences so that the dinucleotide of interest was present in the central position, and then we calculated the GC content of each position across all 2000 sequences. To calculate the width of the peak (or depression) in GC content around dinucleotides we first used the median of the first 1000 GC content values (positions -5000 to -4000 with respect to the central dinucleotide in alignments) as an indicator of the average GC content. In all cases the peak in GC content begins after the first 1000 nucleotides in the alignments and so we believe that this value acts as a good indicator of the average GC content in surrounding sequences. We then assign the start of the peak as the point at which greater than 47 out of 50 base positions in a row have an average GC content across the 2000 sequences that is higher than the median value, labeling the start of the 50bp as the start of the peak. Similarly, the end of the peak is designated in the same way, but running in the opposite direction from the end of the sequences. For a depression in GC content, we required the GC content to be below the median value under the same criteria as before.

3.3.3 Genomic data

Genomic data on telomere and centromere locations, GC content, gene density, nucleosome occupancy and single SNP density were downloaded from the UCSC genome browser (http://genome.ucsc.edu/). Data were downloaded per 1MB for comparison with coincident SNP data. Gene density was taken as the number of base pairs that were part of an exon in each megabase region. A365 values were used for nucleosome occupancy scores as they are comparable to other methods at identifying regions with high nucleosome occupancy (Gupta et al. 2008). Recombination rate data were also averaged across each MB, using data from Kong *et al.* (2002), as were replication-timing scores (S50), which were taken from Chen *et al.* (2010).

We estimated human and chimp divergence per MB as follows. Alignments between the NCBI36 version of the human genome and the PanTro2 version of the chimp genome were downloaded from the UCSC website (http://genome.ucsc.edu/). Nucleotides were masked if they were of low quality in the chimp genome (error rate above 1/10000); quality scores were unavailable for chromosomes 21 and Y and so these were not masked. We then masked any sequences where divergence scores were unexpectedly high; 100bp windows with greater than 10% divergence were masked, with sliding windows every 10bp. Finally, we masked any sequences of less than 20bp that were flanked both sides by >40bp flank sequence. The number of substitutions per MB was calculated in regions containing >100kb of unmasked sequence.

3.3.4 Balancing selection

In order to investigate balancing selection on the PRIM2 gene, and the 175kb region on chromosome 4, we downloaded low-coverage pilot variation data from the 1000 genomes project (http://www.1000genomes.org) that was released in April 2009. The data were split into three groups (CEU, YRI, JPT+CHB) at each locus and comprised of allele frequency data within each population and phased genotype data for each sampled site. We obtained phased haplotypes from the dataset for each region and then used the Neighbour-Joining method in PHYLIP (Felsenstein 2005) to construct a phylogenetic tree within each population. For PRIM2, CEU was sampled across 57 individuals and contained 3974 SNPs, YRI was sampled across 56 individuals and contained 1548 SNPs and the combined populations of JPT+CHB were sampled across 59 individuals and contained 1660 SNPs. For the region on chromosome 4, CEU was sampled across 57 individuals and contained 1070 SNPs, YRI was sampled across 56 individuals and contained 318 SNPs and the combined populations of JPT+CHB were sampled across 59 individuals and contained 284 SNPs.

To calculate the significance of the Tajima's D values we performed a coalescent simulation using MS (Hudson 2002) with the same number of haplotypes as found in each population, assuming a stationary population size and either no recombination or a constant recombination rate that was calculated using the Pairwise program in LDhat (McVean, Awadalla, and Fearnhead 2002). We repeated this procedure 1000 times for each population at each locus and calculated the Tajima's D statistic in each case; the p-value was the number of times the Tajima's D value generated in each of the 1000

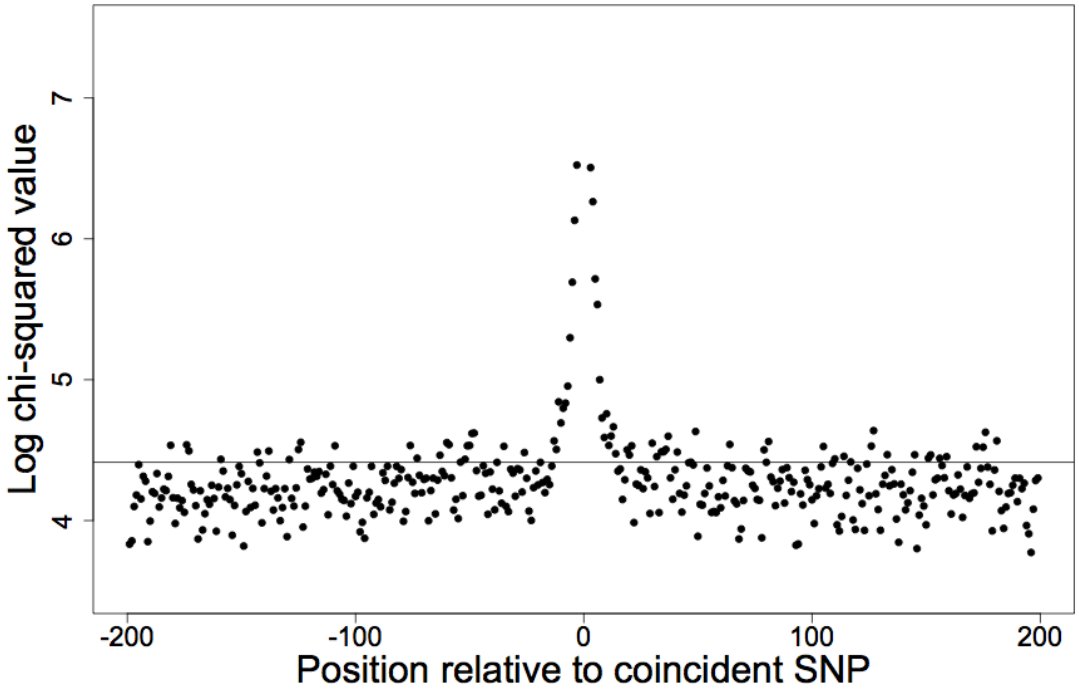
coalescent simulations was greater than the observed value for each population at each locus.

3.4 Results

3.4.1 Local context of coincident SNPs

In the previous chapter we found a significant excess of SNPs in the human genome that also contained a SNP at the orthologous position in chimpanzee; we refer to these as coincident SNPs. Furthermore, we showed that there is significant heterogeneity in triplet frequencies that extends to about 80bp either side of coincident SNPs. We did this by tabulating the frequency of each triplet at each site relative to the coincident SNP across our sequences containing coincident SNPs. The triplet frequencies at each site were then compared to the average frequencies across all sites using a chi-square test. To investigate this pattern further we divided our original data set into CpG and non-CpG coincident SNPs and repeated the analysis as above. For non-CpG coincident SNP sequences (4517 cases) the frequency of triplets within approximately 10bp either side of the coincident SNP are significantly different to the average triplet frequencies across all positions in the sequences (Figure 3.1a). This pattern is entirely driven by runs of A and T nucleotides; if we remove sequences where the coincident SNP falls at the start or the end of a mononucleotide triplet of any kind, the peak in triplet heterogeneity disappears outside of the neighbouring nucleotides (Figure 3.1b). It should be noted at this point that mononucleotide runs are not the cause of the excess of coincident SNPs, since removing SNPs that form part of a run of 3 or more nucleotides still leaves a large

a)



b)

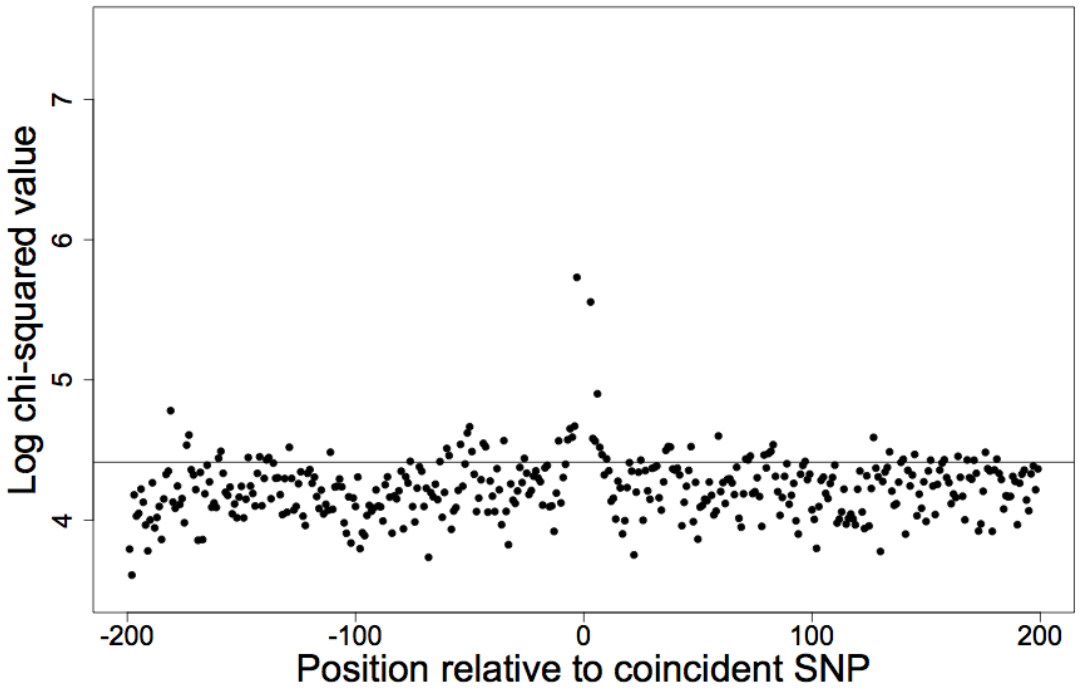


Figure 3.1. Heterogeneity in triplet frequencies around non-CpG coincident SNPs. Figure gives the log chi-squared value of the heterogeneity of triplet frequencies against the average triplet frequencies across the whole alignment for (a) all alignments containing a non-CpG coincident SNP and (b) alignments where the SNP is not part of a mononucleotide run of 3 or more nucleotides. The horizontal line marks the 5% significance value for the chi-square test.

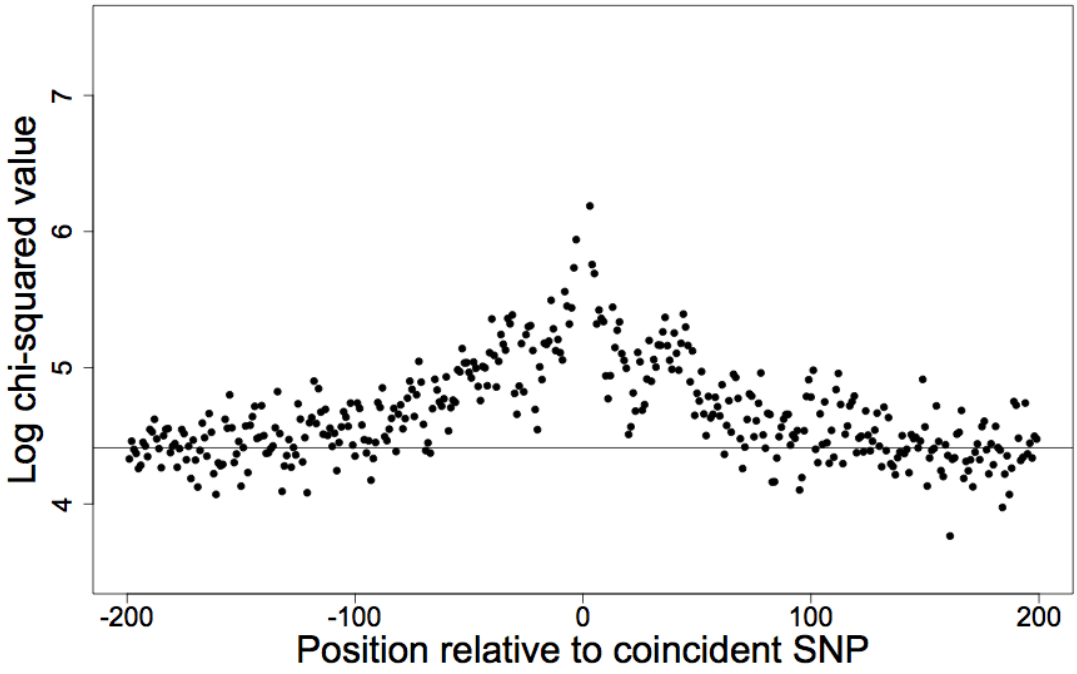
excess of coincident SNPs (see chapter 2). There therefore appears to be no heterogeneity in triplet frequencies around non-CpG coincident SNPs when mono-nucleotide runs are removed.

For CpG coincident SNP sequences (5930 cases) the heterogeneity of triplet frequencies extends up to ~100bp either side of the coincident SNP (Figure 3.2a). However, if we take the same number of CpGs, which do not contain a SNP, from unique sequences at random from the human genome we observe a very similar pattern (Figure 3.2b). This indicates that the pattern around CpG coincident SNPs is entirely dominated by the pattern around CpG dinucleotides, whether they contain a SNP or not, and thus there are no local context effects associated specifically with coincident SNPs. As such, variation in the mutation rate in the human genome that was inferred from the excess of coincident SNPs is truly cryptic on a local scale, in the sense that there do not appear to be any non-random patterns of nucleotides in the surrounding sequence.

3.4.2 Patterns around CpG and other dinucleotides

Although we have shown that there are no local nucleotide contexts associated specifically with coincident SNPs, it is interesting to consider what is driving the patterns in triplet heterogeneity around CpG dincleotides that do not contain a SNP. In order to investigate this we estimated the frequencies of triplets we would expect to see given the single nucleotide compositions at each position in the sequences containing CpG dinucleotides. Interestingly, we found that the peak all but disappeared (Figure 3.3a). Furthermore, if we plot the GC content across alignments at each position, there is a similar peak around the central CpG dinucleotide (Figure 3.3b). This implies that

a)



b)

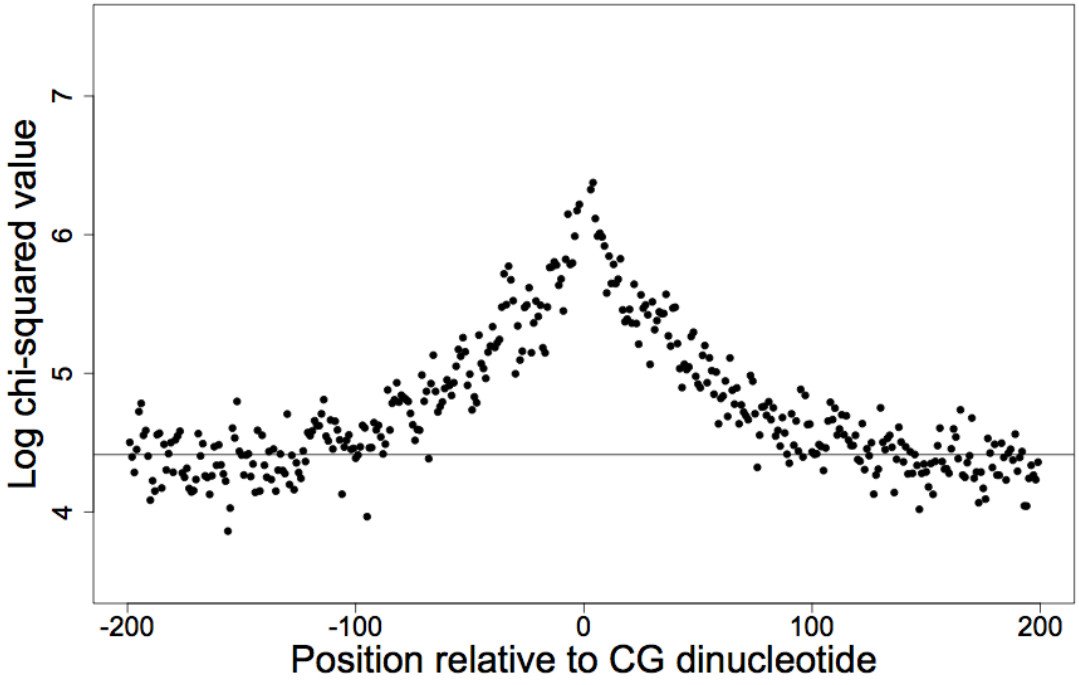
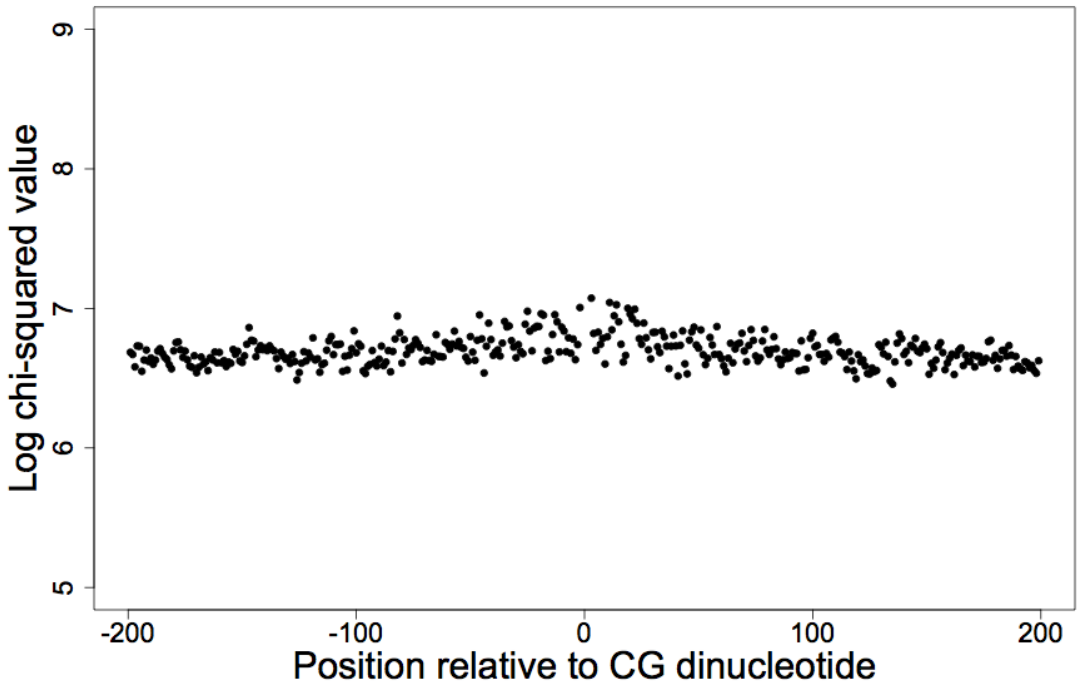


Figure 3.2. Heterogeneity in triplet frequencies. Figure gives the log chi-squared value of the heterogeneity of triplet frequencies against the average triplet frequencies across the whole alignment for (a) all alignments containing a CpG coincident SNP and (b) sequences that contain a CpG dinucleotide but no SNP. The horizontal line marks the 5% significance value for the chi-square test.

a)



b)

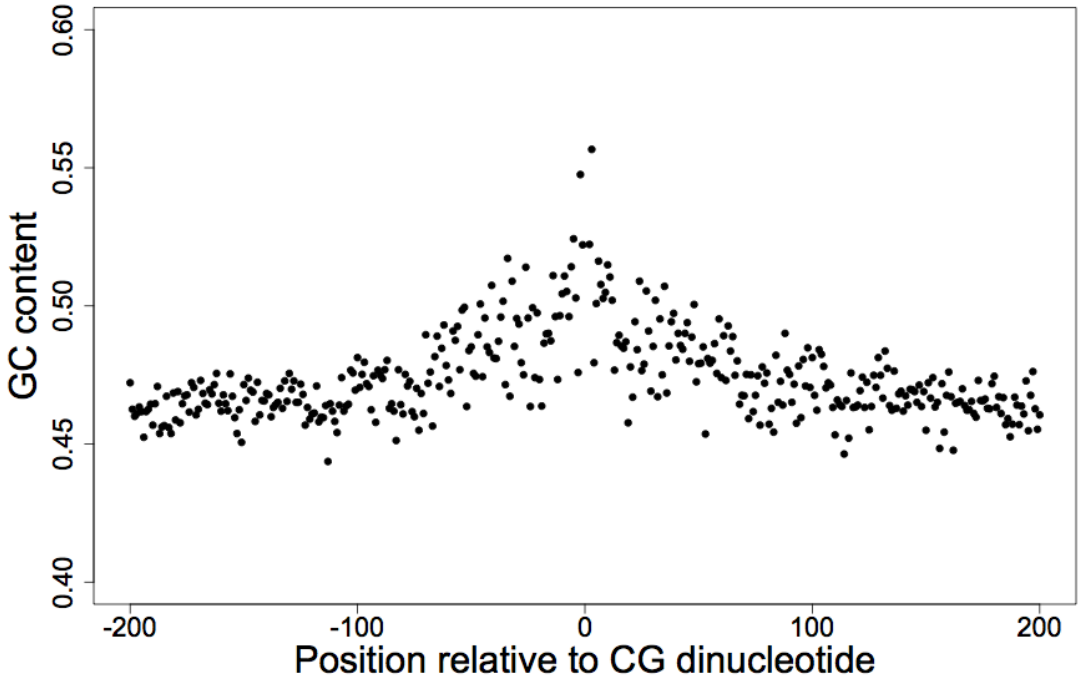


Figure 3.3. Nucleotide patterns around CpG dinucletides that do not contain a SNP. Figure (a) gives log chi-squared values across alignments when single nucleotide frequencies are used to predict triplet frequencies at each position and (b) shows the GC content at each position in the alignment.

there is a general increase in GC content around CpG dincleotides in the human genome, a pattern has also been previously observed in a study by Elango *et al* (2008), and that this explains the pattern of triplet heterogeneity in regions around CpGs.

The question therefore arises as to whether there are patterns in nucleotide content around other non-SNP containing dinucleotides in the human genome. To answer this, we selected 2000 cases of each type of dinucleotide at random from the human genome, flanked by 5000bp either side. We then lined up the sequences so that the nucleotide of interest was present in the central positions and then considered the GC content across the sequences at each position. Sequences were only considered if they did not contain possible CpG islands (see methods). However, as it is likely that some dinucleotides are part of mono-nucleotide and dinucleotide runs, we selected only dinucleotides that were not part of two or more of the same dinucleotides or where the first or second base of the dinucleotide was not part of a mononucleotide triplet. As expected, we find an increase in GC content around CpG dinucletides that runs from -231bp to 199bp (with the dinucleotide at position zero), but also a peak in GC content around GpC dinucleotides that runs from -77bp to 90bp, which are shown in figures 3.4a and 3.4b respectively. Furthermore, there is a decrease in GC content around TA dinucleotides that extends from -95bp to 80bp (Figure 3.4c). The specific widths of the peaks are clearly determined by the number of sequences used, however they are useful in comparisons of different dinucleotides, and all clearly show a context effect. Consequently, it appears that there are strong nucleotide patterns acting on a very local scale in the human genome. There are also peaks in GC content around GGs and CCs, and troughs in GC content around AAs and TT, however for GGs and CCs this is caused by sequences in which a CpG or GpC is found immediately adjacent to the

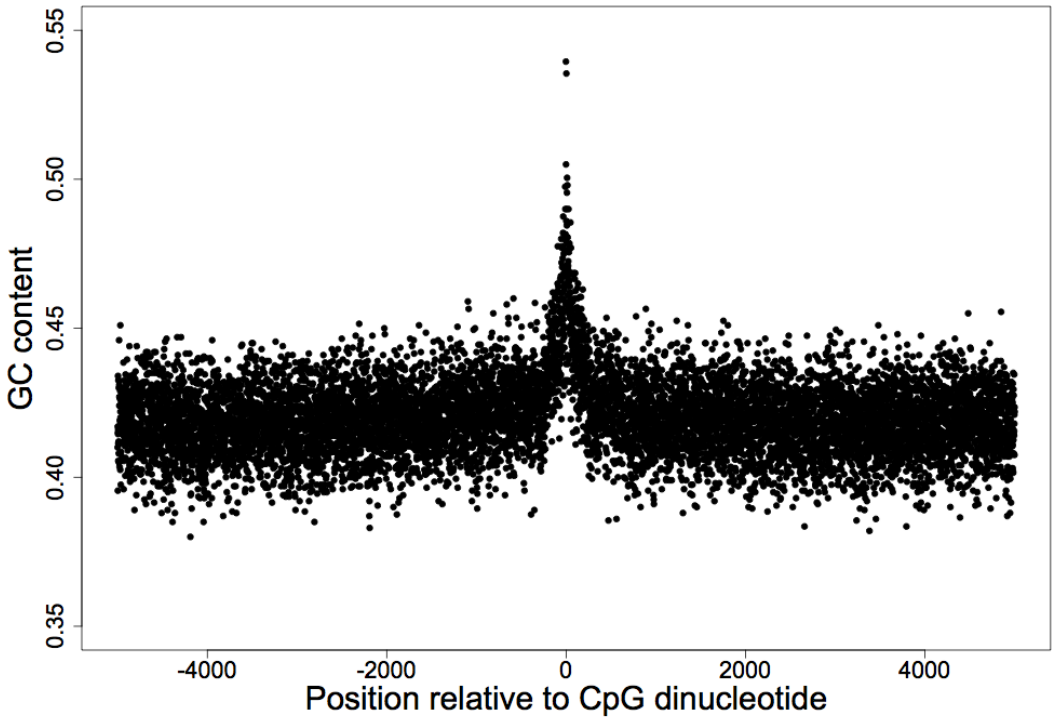


Figure 3.4a. GC content around CpG dinucleotides

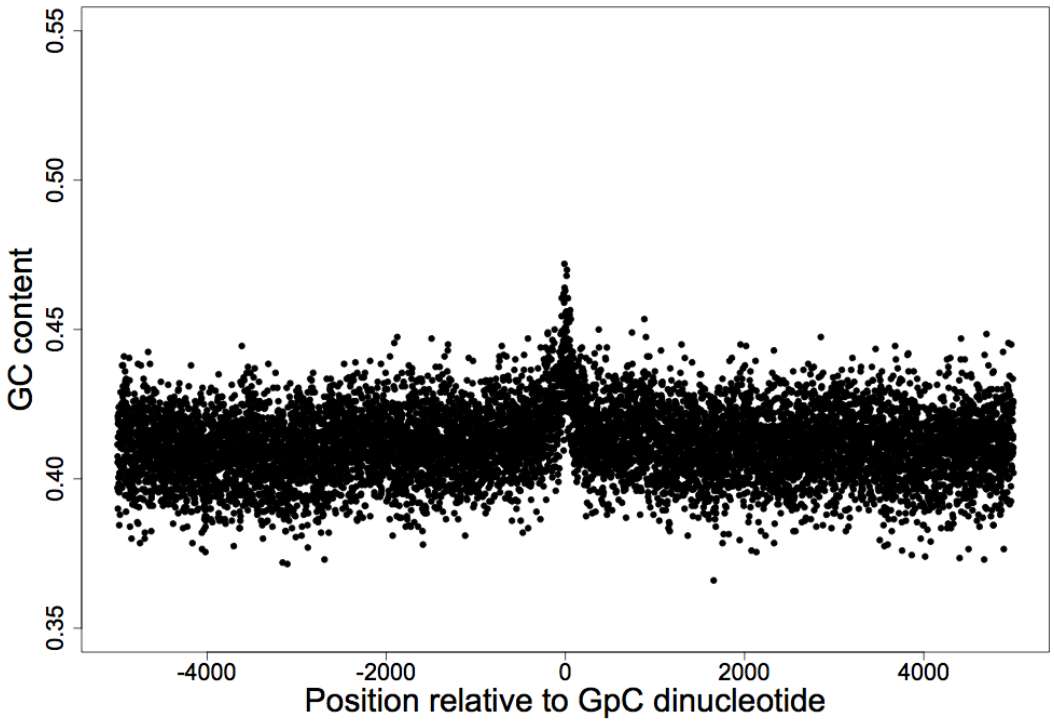


Figure 3.4b. GC content around GpC dinucleotides.

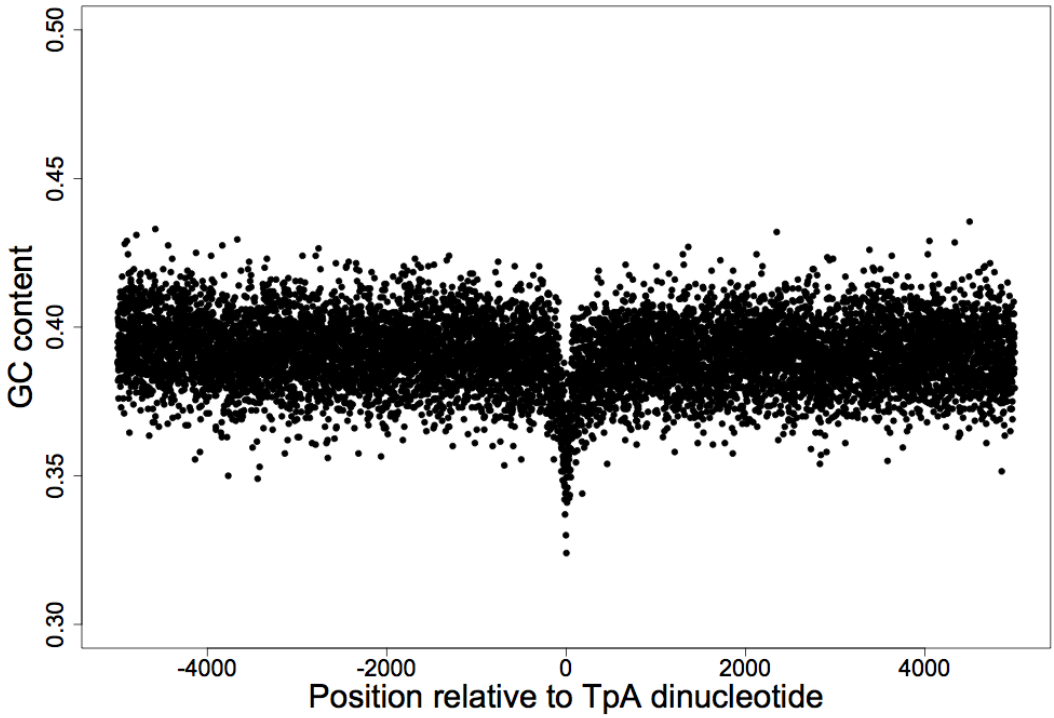


Figure 3.4c. GC content around TpA dinucleotides.

central dincleotide, and for AAs and TTs where TA or AT is found immediately adjacent to the central dinucleotide. If these sequences are removed, the patterns disappear. There are no patterns in GC content around other dinucleotides.

3.4.3 Genomic distribution of coincident SNPs

To investigate the genomic distribution of coincident SNPs we split the human genome into regions of 1MB and tallied the number of coincident and single (non-coincident) SNPs found in each region. Regions with no SNPs were excluded from further analysis; these are typically found in the heterochromatic regions near the centromere. On average there were 8014 and 3.91 simple and coincident SNPs per MB respectively, 6838 and 1.68 of which were non-CpG. If coincident SNPs occurred at random across the human genome, then we would expect the number of coincident SNPs in each 1MB region to be Poisson distributed and to have a variance equal to the mean. However, the observed variance, 13.27, is far in excess of this, and using a chi-square test we find that the number of coincident SNPs per MB is significantly overdispersed ($p<0.001$); for example the third quartile is 2.5 fold higher than the first quartile, whereas it would be expected to be 1.67 fold higher if coincident SNPs were distributed at random. Therefore, coincident SNPs are non-uniformly distributed across the human genome. It is possible that the distribution of coincident SNPs is not Poisson distributed across the human genome if there is variation in the level of sampling that has occurred between different regions. However, this would lead to an excess of coincident SNPs through general clustering of single SNPs, which we do not observe when considering the

distribution of distances between human and chimpanzee SNPs in the original analysis (see chapter 2).

The distribution of single SNPs is also known to be non-uniform (Venter et al. 2001), and we find that the density of coincident SNPs has a significantly positive correlation with the density of single SNPs ($r^2 = 0.065$, $p<0.001$ for all SNPs and $r^2 = 0.037$, $p<0.001$ for non-CpG SNPs); this is perhaps not surprising given that SNP densities must drive the locations of coincident SNPs to a certain extent, since at least half of coincident SNPs are thought to be due to chance alone, as single SNPs coincide at random (chapter 2). However, the correlation between the density of single and coincident SNPs is not strong, and the lack of a strong correlation is not due to high sampling error in coincident SNP density; as we have shown above, the observed variance in the density of coincident SNPs is substantially greater than we expect from sampling error alone – i.e. the distribution of coincident SNP density is overdispersed. We can estimate the approximate proportion of the variance in coincident SNP density that is due to sampling error as follows; since we expect the number of coincident SNPs in each genomic region to be Poisson distributed, the average sampling variance is likely to be of the order of the mean number of coincident SNPs per MB; this is approximately 30% of the total variance in the density of coincident SNPs. Given that the correlation between the density of single SNPs only explains 6.5% of the density in coincident SNPs, it is evident that the poor correlation is not due to sampling error in the density of coincident SNPs.

To investigate the variation in coincident SNP density in more depth we compared the frequency of coincident SNPs in each 1MB to some key genomic features (table 3.1).

There is no significant correlation between the density of coincident SNPs and the distance to the centromere, the distance to the nearest telomere or the nucleosome association, the gene density, replication timing or GC content of a region. There is, however, a significantly positive correlation between coincident SNP density and recombination rate ($r = 0.107$, $p<0.001$). This may reflect the significant correlation that exists between the density of single SNPs and the rate of recombination ((Hellmann et al. 2003; Hellmann et al. 2005); in our dataset $r =0.234$, $p<0.001$), and the fact that approximately half of all coincident SNPs appear to be due to chance alone. To investigate this further we performed a partial correlation of coincident SNP density against the rate of recombination controlling for the single SNP density and found that the correlation is still significantly positive ($r = 0.048$, $p=0.011$). However, despite a significant correlation, very little variation in coincident SNP density is explained by recombination rates as the correlation has a very low r^2 value of 0.002. It is also interesting to note that there are significant correlations between single SNP densities and the same set of genomic features as mentioned above (table 3.2), suggesting that there are subtle differences between the distributions of coincident and single SNPs in the human genome.

It is puzzling that the density of single SNPs does not significantly correlate with replication timing, since it has been previously shown that primate divergence rates are higher in late replicating regions, suggesting that they have a higher mutation rate (Stamatoyannopoulos et al. 2009; Chen et al. 2010). Furthermore, Stamatoyannopoulos *et al.* (2009) showed that replication also correlates with SNP density over a scale of 100kb, although different SNP and replication timing datasets were used in this analysis. In an attempt to explain this discrepancy we divided human SNP density by

Feature	r	p
SNP Density	0.256	<0.001
Distance to Telomere	-0.022	0.226
Distance to Centromere	0.011	0.565
Recombination Rate	0.107	<0.001
Nucleosome Association	0.004	0.832
Gene Density	-0.022	0.230
GC Content	-0.006	0.741
Replication Timing	0.004	0.838

Table 3.1: The correlation between the number of coincident SNPs per MB and various genomic features.

Feature	r	p
Distance to Telomere	-0.171	<0.001
Distance to Centromere	-0.047	0.012
Recombination Rate	0.234	<0.001
Nucleosome Association	0.187	<0.001
Gene Density	0.064	0.001
GC Content	0.184	<0.001
Replication Timing	0.008	0.673

Table 3.2: The correlation between the number of single SNPs per MB and various genomic features.

the average divergence between human and chimp per MB to give an estimation of the effective population size (N_e) for each region and compared this to replication timing; we find a significant negative correlation ($r=-0.200$, $p<0.001$). We then performed a partial correlation between single SNPs and replication timing, whilst controlling for N_e, and we observe a significant positive correlation ($r=0.276$, $p<0.001$), suggesting that a negative relationship between N_e and replication timing density may be canceling out a relationship between diversity and replication timing. Furthermore, if we perform a partial correlation between coincident SNP density and replication timing, whilst controlling for N_e, we observe a significant positive correlation ($r=0.102$, $p<0.001$), although replication timing explains very little of the variance in the distribution of coincident SNPs (~1%).

3.4.4 Regions in the human genome with high numbers of coincident SNPs

There are two 1MB regions in the human genome that contain considerably more coincident SNPs than any other region and are outliers in the distributions of both all and non-CpG coincident SNPs. These regions are chromosome 4, 190mb - 191mb, which contains 57 coincident SNPs, and chromosome 6, 57mb - 58mb, which contains 100 coincident SNPs. The region on chromosome 4 falls very close to the end of the chromosome and contains no known genes, however 53 coincident SNPs are found in the region running from 190712230 to 190887438 (www.hapmap.org), which is approximately 175kb in length. A gene called PRIM2, which codes for the large DNA primase subunit, dominates the region on chromosome 6 (Shiratori et al. 1995). The smaller primase subunit is encoded by PRIM1 and DNA primase is a polymerase that plays an important role in DNA replication by synthesizing small RNA primers, which

can then be used as starting points for the production of Okasaki fragments by the lagging strand polymerase during discontinuous replication (Roth 1987). Primase also acts to prevent leading strand synthesis outpacing lagging strand synthesis by acting as a transient brake by halting the progress of the replication fork (Lee et al. 2006). PRIM2 is located at 57,290,381-57,621,334 (www.hapmap.org) and contains 86 coincident SNPs, all of which are intronic. The high number of coincident SNPs in PRIM2 and near the telomere of chromosome 4 could be due to a concentration of cryptically hypermutable sites, however they could also be due to long-term balancing selection maintaining polymorphisms between species. It is also possible that the high concentration of coincident SNPs are a result of segmental duplications that increase the chances of chimpanzee SNPs being coincident with human SNPs if there are two or more almost identical regions on the human genome that the sequences surrounding SNPs could match to. It is important to note that this is not the cause of the significant excess of coincident SNPs in general, since we showed in chapter 2 that paralogous sequences do not contribute substantially to the excess of human and chimpanzee coincident SNPs. Similarly, balancing selection cannot explain the excess of coincident SNPs across the whole genome, since we also showed that there is an excess of coincident SNPs between human and macaque; the large divergence between the two species would mean that balancing selection would be extremely unlikely (see chapter 2).

To investigate the matter further we downloaded SNP data from the 1000 genome project (http://www.1000genomes.org) for the three Hapmap populations that have already been sequenced. The region in the PRIM2 locus, which has a high concentration of coincident SNPs, also has a relatively high density of single SNPs in

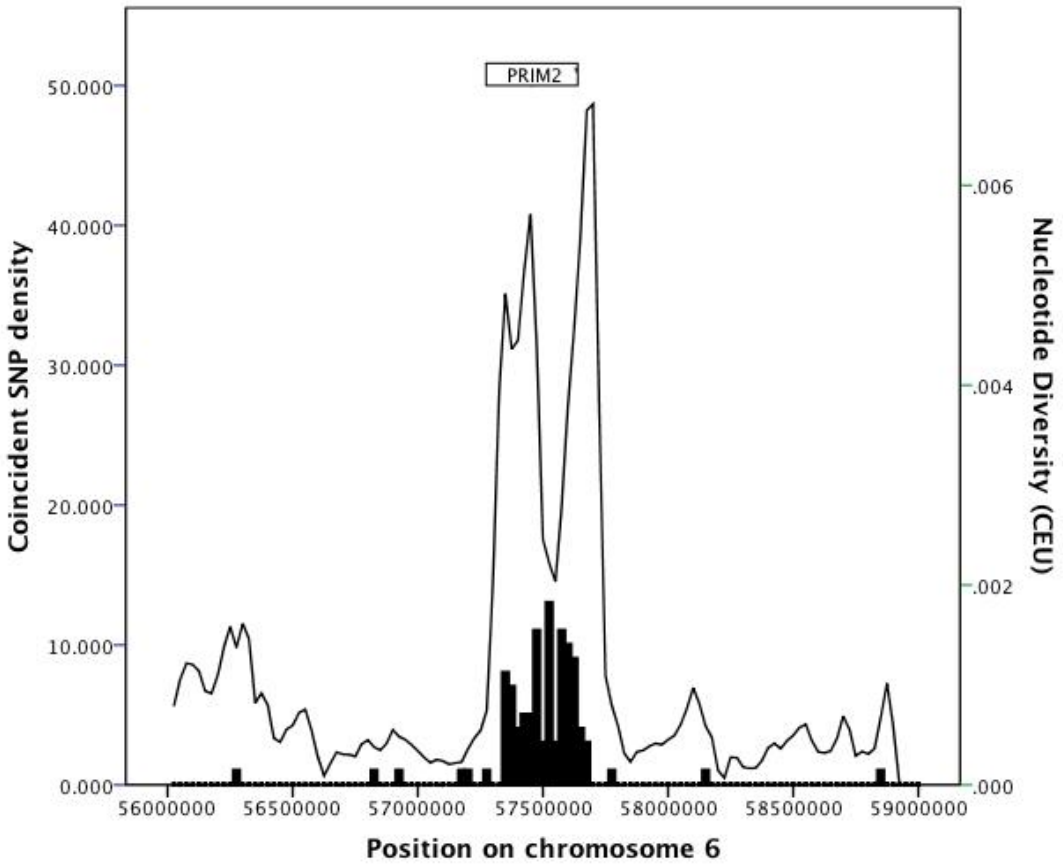


Figure 3.5. The nucleotide diversity across the region containing PRIM2 for the CEU population. The figure shows a sliding window of nucleotide diversity every 25kb, with window size of 50kb as a line graph corresponding to the right hand axis, with the coincident SNP density as a bar chart corresponding to the left hand axis.

all three human populations (figure 3.5 for the European population, appendix 3.1 for other populations); however the region with the very highest density of coincident SNPs has a relatively low density of single SNPs within the PRIM2 locus. Tajima's D is significantly positive in the PRIM2 locus in all three human populations (D = 2.310 in Europeans, 1.296 in Yorubans, and 1.253 in East Asians, $p<0.0001$ assuming a constant rate of recombination). It could be argued that since SNP calling is conservative, many rare variants will have been missed, possibly leading to an artificially high Tajima's D statistic for the region. However, when we perform a sliding window analysis, calculating the Tajima's D statistic in each window of width 200kb every 100kb (overlapping windows), we clearly see a peak in the statistic above that observed in surrounding regions (figure 3.6 for the European population, appendix 3.2 for other populations). There might therefore be some evidence of balancing selection acting in this locus, particularly in the European population. However, under balancing selection we might expect to see groups of divergent haplotypes, and this is not what we observe if we construct phylogenetic trees of inferred haplotypes in each population (results not shown).

In contrast, the region on chromosome 4 with a high concentration of coincident SNPs has one of the lowest densities of single SNPs in the region, especially for the African and East Asian populations (figure 3.7 for the Yoruba population, appendix 3.3 for the other populations). Furthermore, Tajima's D is 1.654 for the European population, 0.827 for the East Asian population and 0.647 for the Yoruban population in the region with the highest concentration of coincident SNPs, running from 190712230 to 190887438 on chromosome 4. Although the Tajima's D scores are all significantly positive assuming a constant rate of recombination ($p<0.01$ for all populations), only the

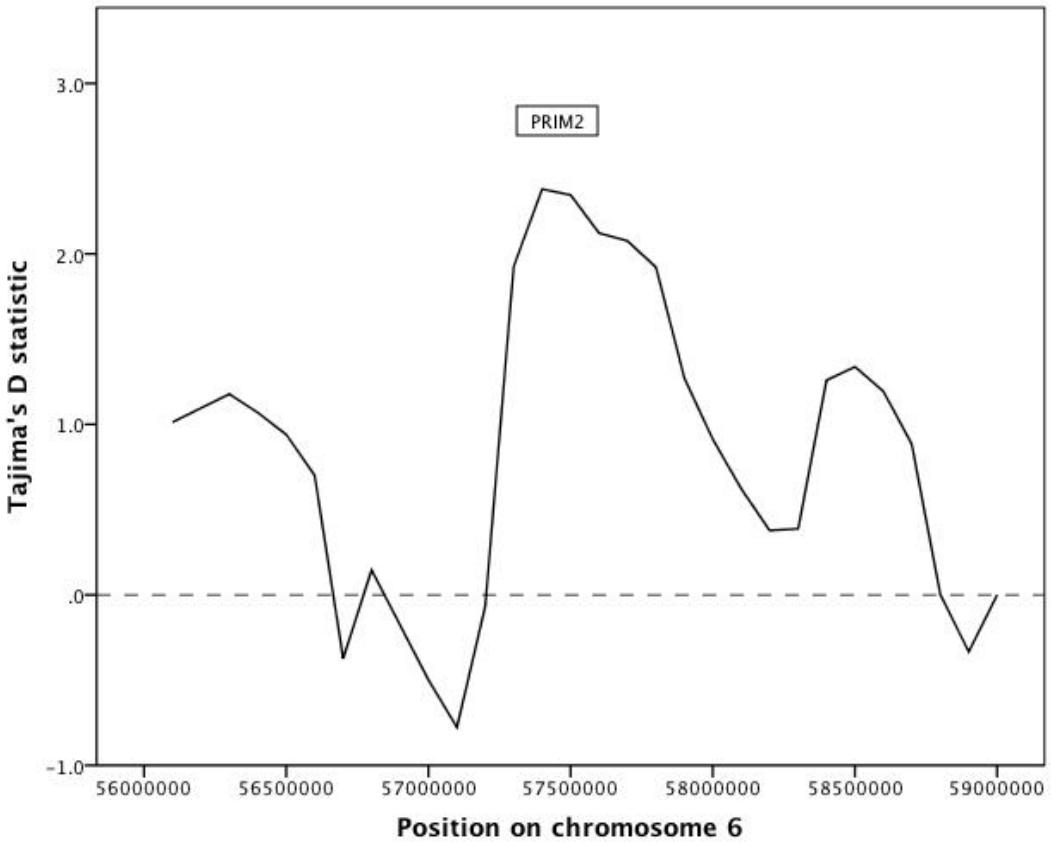


Figure 3.6. The Tajima's D statistic across the region containing PRIM2 for the CEU population. The figure shows a sliding window of nucleotide diversity every 100kb, with window size of 200kb.

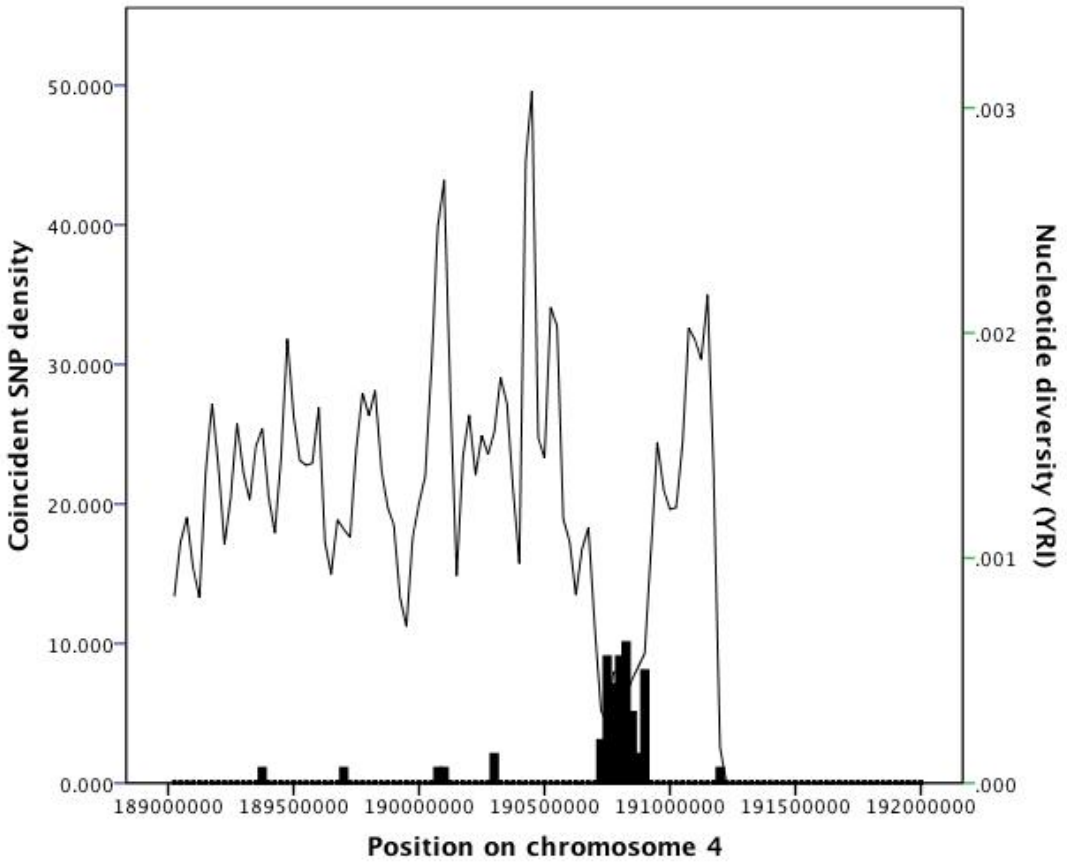


Figure 3.7. The nucleotide diversity across the region on chromosome 4 for the YRI population. The figure shows a sliding window of nucleotide diversity every 25kb, with window size of 50kb as a line graph corresponding to the right hand axis, with the coincident SNP density as a bar chart corresponding to the left hand axis.

European population has a Tajima's D in excess of one and in all three populations, the Tajima's D statistic is considerably lower than those observed at PRIM2, possibly making PRIM2 a more likely candidate for balancing selection. Furthermore, as with PRIM2, if we construct phylogenetic trees of inferred haplotypes in each population, we do not see distinct groups of divergent haplotypes (results not shown).

The PRIM2 region is not part of any known segmental duplications. However, for the 175kb region on chromosome 4 that contains a high density of coincident SNPs, there are two regions that have undergone segmental duplication which contain a total of 20 coincident SNPs and span approximately 70kb (Bailey et al. 2001; Bailey et al. 2002). The average sequence identity between these two regions and their corresponding duplications is approximately 96.56%. Under our criteria for detecting coincident SNPs in homologous sequences between human and chimpanzee (requiring a match of at least 96 out of 101 sites for alignments containing coincident SNPs) and assuming that the number of mismatches per alignment is Poisson distributed, we expect that on average 46% of the chimpanzee sequences would not match to the associated duplicated region in humans. This allows us to conclude that approximately 42 of the 53 coincident SNPs found in the 175kb region are correctly matched, which is still considerably above average, and therefore that segmental duplications are not having a great impact on the numbers of coincident SNPs in this region.

3.5 Discussion

In the previous chapter we provided evidence for cryptic variation in the mutation rate in the human genome by showing that there is a significant excess of SNPs that are present at orthologous positions in human and chimp. We also suggested that there are complex context effects associated with coincident SNPs by showing there was significant heterogeneity in triplet frequencies up to ~80bp either side of coincident SNPs. However, we show here that this pattern is not specific to coincident SNPs and is instead present around CpG dinucleotides, regardless of whether the dinucleotide contains a SNP. To that end, it seems that cryptic variation in the mutation rate is not dependent upon specific local context effects that are associated with coincident SNPs and that the general excess of coincident SNPs is not driven by local context effects; indeed no single context can explain more than a very small fraction of coincident SNPs.

The distribution of coincident SNPs across the genome is non-uniform such that some parts of the genome have higher densities of SNPs than others. However, this variation is not strongly correlated to any obvious genomic feature. There are a small number of regions that have a very high number of coincident SNPs, and we studied these in more detail to investigate whether balancing selection might be involved. We obtained inconclusive results. Although the PRIM2 region has a high nucleotide diversity and Tajima's D is significantly positive, the region with the very highest density of coincident SNPs did not have a particularly high diversity. Furthermore, a phylogenetic tree of haplotypes did not reveal evidence for any deep branches, which might be indicative of long term balancing selection. The region on chromosome 4 with a high

density of coincident SNPs actually has rather low diversity for the genomic region in which it resides. Other studies have shown that genome wide testing for balancing selection has thus far been fruitless, leading to a suggestion that it is either rare or hard to identify (Bubb et al. 2006). Interestingly, the MHC locus, a region that is thought to be undergoing strong balancing selection (Hughes and Nei 1988), has an average of 11.4 coincident SNPs per MB, which is far above the genome-wide average of 3.91, but markedly below the densities found at PRIM2 and the 175kb region on chromosome 4.

Finally, the patterns around CpG dinucletides are driven almost entirely by GC content, which increases from ~200bp either side of the dinucleotide up to a peak immediately adjacent to the dinucleotide. This distance-decaying pattern is almost identical to that seen in a study by Elango *et al.* (2008). We also observe a similar pattern in sequences surrounding GpC dinucleotides, albeit to a smaller extent, whilst the inverse is true of TpA dinucleotides, with the GC content decreasing towards the two central bases. The pattern in GpC dinucleotides is also evident in the work of Elango *et al.* (2008); however they did not draw attention to this pattern since it operates over a limited scale and they were interested in larger scale processes. It seems likely that the increase in local GC content around CpGs is caused by a process suggested by Fryxell and Zuckerkandl (2000) in which CpG dinucleotides mutate more rapidly in AT rich regions due to an increased rate of DNA duplex melting. Cytosine deamination occurs ~143 times more often on ssDNA than it does on dsDNA (Frederico, Kunkel, and Shaw 1990), thus the mutability of CpGs is closely linked to the melting temperature of the surrounding DNA. A 10% decrease in GC content reduces the melting temperature of a sequence by 4.1^{o}C, and thus increases the deamination of methylated cytosine by 2-fold (Fryxell and Zuckerkandl 2000). This process could also explain the increase in GC

content around GpCs, as Fryxell and Moon (2005) note that unmethylated GpCs undergo deamination at lower rates in GC rich regions due to reduced DNA melting. Furthermore, we may expect TpA dinucleotides to be present in AT-rich regions if they are remnants of former CpGs that have mutated at high rates in the past. An alternative explanation is that biased gene conversion (BGC) alters the base composition around CpG dincleotides (Elango et al. 2008). BGC is a mechanism in which base mismatches formed during recombination and the repair of double strand breaks in heterozygous individuals are preferentially repaired to GC over AT nucleotides (Marais 2003). This process may decrease the mutability of CpG dinucleotides near to recombination events if deaminated cytosines are preferentially repaired or if CpGs end up in less mutable GC rich contexts as a result of BGC. However, this does not readily explain why the local increase in GC content is so much more conspicuous for CpG as opposed to GpC dinucleotides, and why patterns are absent around CC and GG dinucleotides. Furthermore, it is not obvious how BGC could generate a decrease in GC content around TpA dinucleotides, as regions between areas undergoing high levels of recombination, in which A and T nucleotides are relatively more likely to accumulate, are likely to be much larger than the ~200bp over which the pattern appears to exist. In order to differentiate between the two potential mechanisms more formally we tested for the presence of peaks in GC content around CpGs in regions that do not undergo recombination on the Y chromosome in *H. Sapiens*. As there is no recombination in these regions, and thus no BGC, we should see no differences in base composition across the sequences if the patterns around CpGs are caused by BGC. We repeated the procedure outlined above by obtaining 2000 sequences from the Y chromosome and calculating the GC content at each position across the alignments. It has been reported that certain regions of the human Y chromosome undergo gene conversion (Skaletsky et

al. 2003) and so these regions were not included in the analysis. For sequences containing CpG dinucleotides from the Y chromosome in *H. sapiens* there is a peak of GC content that extends from -160bp to 186bp, showing that the pattern exists independently of recombination. It therefore seems that the melting temperature of different sequences can drive local biases in base composition around certain dinucleotides in the human genome.

We have shown that there are no obvious sequence contexts surrounding coincident SNPs, which we have inferred to be caused by cryptic variation in the mutation rate. Furthermore, we have failed to find any genomic feature that correlates strongly to the density of coincident SNPs. What then might cause some sites to have much higher mutation rates than others? It seems likely that it is caused by DNA topology and packaging, but until we understand these processes in the germ-line we may struggle to understand this phenomenon further.

Acknowledgements. AH and AEW were funded by the BBSRC. We thank two anonymous referees for comments.

4. Human tri-allelic sites: evidence for a novel form of mutation?

4.1 Abstract

Most SNPs in the human genome are bi-allelic, however there are some sites that are tri-allelic. We show here that there are approximately twice as many tri-allelic sites as we would expect by chance. This excess does not appear to be caused by sequencing errors, natural selection or mutational hotspots that result in a single mutation. Instead, we propose that the excess of tri-allelic sites is driven by new mutations inducing another mutation either within the same individual, or subsequently during recombination. We provide evidence for this model by showing that the rarer two alleles at tri-allelic sites tend to cluster on phylogenetic trees of human haplotypes. However, we find no association between the density of tri-allelic sites and the rate of recombination, which leads us to suggest that tri-allelic sites might be generated by the simultaneous production of new two mutations within the same individual on the same genetic background. Under this model we estimate that simultaneous mutation contributes approximately 3% of all distinct SNPs. We also show that there is a two-fold excess of adjacent SNPs. Approximately half of these seem to be generated simultaneously since they have identical minor allele frequencies. We estimate that the mutation of adjacent nucleotides accounts for a little less than 1% of all SNPs.

4.2 Introduction

Although the density of bi-allelic SNPs in the human genome is reasonably low, there are some sites that have three (tri-allelic sites) or even four nucleotides segregating in the human population. We show here that there are approximately twice as many tri-allelic sites as we would expect by chance. There are at least three types of mutational mechanisms that could potentially generate such an excess of tri-allelic sites. First, some sites may be hypermutable, and if the mutation rate of at least two pathways (e.g. C->T and C->A) is elevated at such sites, then there will be an excess of tri-allelic sites. The mutation rate of a site is known to depend upon the adjacent nucleotides, the best known example being the CpG dinucleotide (Coulondre et al. 1978; Bird 1980) at which both the frequency of transition and transversion mutations is elevated. However, other adjacent nucleotides also influence the mutation rate (Blake, Hess, and Nicholsontuell 1992; Zhao et al. 2003; Hwang and Green 2004). Furthermore, we showed in chapters 2 and 3 that there is variation in the mutation rate that does not dependent upon the identity of the adjacent nucleotides, or any specific context.

Second, it is possible that two of the alleles at a tri-allelic site are generated simultaneously within a single individual. Point mutations are generally assumed to involve the production of a single new allele per mutation event at a rate of which is governed by the effects mentioned above. However, it is not difficult to imagine mechanisms that might induce mutations on both strands of the DNA duplex; for example, the presence of a base mismatch may itself be unstable, so we might go from a G-C base pair to a G-A, which then may mutate to C-A; if DNA replication reads through this mismatch, the G allele will have mutated to both C and T. Alternatively,

the mutation may occur across both strands of the duplex at the same time, possibly as a result of a chemical or radiation event. Thirdly, in a similar manner, we might imagine a single SNP inducing subsequent mutations if base mismatches are formed during recombination in heteroduplex DNA.

Here we attempt to identify the process causing the excess of tri-allelic sites by analysing sequence data around tri-allelic sites.

4.3 Materials and Methods

4.3.1 Tri-allelic sites

The expected number of tri-allelic sites in nuclear DNA was estimated as follows. We downloaded human SNP data from the Environmental Genome Project (NIEHS SNPs 2008) and the SeattleSNPs project (SeattleSNPs 2008). High quality sequence data were used to identify SNPs (Q > 25), and each SNP reported was confirmed in multiple individuals and/or multiple reactions. Assuming the same Q value for re-sequencing, the error rate is $1x10^{-5}$. We masked all CpG and coding sites; coding sites were removed since it is difficult to calculate the expected number of tri-allelic sites in coding sequences because of selection. Sites were designated as CpG if the site, or any of the SNPs at the site, would yield a CpG dinucleotide. We started by tallying the number of each type of nucleotide within each intron and across all genes, ignoring any regions that were not scanned for variation. We then calculated the frequency of each type of SNP, μ_{X-Y} where X and Y are either A,C,T or G, by orientating the SNP using

orthologous chimpanzee sequences and then dividing the number of human sites where X was inferred to be the ancestral allele and Y was inferred to be the derived state by the total number of X sites. For example, μ_{A-G} was estimated by dividing the number of sites where the inferred mutation was from A to G (i.e. A was the allele present at the orthologous chimpanzee position and G was the second allele at the SNP site) by the total number of sites that were A. Orthologous chimpanzee sequences were found and downloaded using Ensembl Biomart (http://www.ensembl.org/biomart/martview/), and then aligned to human sequences using FSA (http://math.berkeley.edu/~rbradley/papers/manual.pdf), which incorporates Exonerate (Slater and Birney 2005) and MUMmer (Kurtz et al. 2004). We were unable to find a small number of orthologous chimpanzee sequences and there were occasional gaps in alignments. In total, ~94% of human SNP sites had an orthologous nucleotide in chimpanzee; the ancestral state was inferred from the major allele at sites with no orthologous chimp nucleotide. At tri-allelic sites, two mutations were assumed to have occurred. The expected number of tri-allelic sites (E_{mt}) was then found by multiplying each mutation rate by the total frequency of nucleotides with the same allele ($n(X)$):

$$E_{mt} = \sum n(X) \,.((\mu_{X-Y} \,.\, \mu_{X-Z1}) + (\mu_{X-Y} \,.\, \mu_{X-Z2}) + (\mu_{X-Z1} \,.\, \mu_{X-Z2})) \qquad (4.1)$$

where the summation is across X, and Z1 and Z2 are either A,C,T or G, and X, Y, Z1 and Z2 are all different nucleotides in each case.

We also downloaded >5000 complete human mitochondrial sequences from GenBank (http://www.ncbi.nlm.nih.gov) and aligned the protein coding sequences. Sequences in which genes were of different length to the consensus were removed, leaving 4764

complete alignments. We considered four-fold synonymous sites only. The expected number of tri-allelic sites in mitochondrial sequences was found in much the same way, using the same formula as that used for nuclear DNA (equation 4.1). However, in this case the ancestral state was inferred from the major allele at each site, as orientation of SNPs using the chimp sequence is impossible because of the large divergence between humans and chimps for mtDNA. Although the use of frequency data will lead to some level of misinference, this is likely to be very small because when the population size is stationary the level of misinference is expected to be ~7% for 5000 sequences. mtDNA also shows a skew towards rare alleles, which will further reduce the level of misinference.

4.3.2 Distribution of single SNPs

The expected distances between SNPs were estimated by randomly distributing SNPs within each intron across the intron sequence. SNPs were not allowed to fall on CpG dinucleotides, as these would have been discarded as CpG SNPs (as stated above). The expected number of tri-allelic sites was the number of times a site was hit by two SNPs multiplied by a factor K, which reflects the fact that when two mutations occur at the same site they will not necessarily generate a tri-allelic SNP; for example, if two transitions occur at the same site. Let the proportion of SNPs that are transitions be f_{ts} and the proportion of transversions that are G<->T or C<->A be f_{tv1} and the proportion that are G<->C or A<->T be f_{tv2}. Then, the expected number of tri-allelic sites is $2x^2 . f_{ts} (f_{tv1} + f_{tv2}) + 2x^2 . f_{tv1} . f_{tv2}$, where x is the density of SNPs, and the expected number of times two SNPs are expected to fall at the same site is x^2. Thus $K = 2(f_{ts} (f_{tv1} + f_{tv2}) + (2$

f_{tv1} . f_{tv2})). In the human genome f_{ts} = 0.66, f_{tv1} = 0.183 and f_{tv2} = 0.163 (http://www.ncbi.nlm.nih.gov/), which means that K=0.516.

The expected number of tri-allelic SNPs incorporating the effects of adjacent nucleotides on the rate of mutation was calculated using equation 4.1, but by summing X over triplets rather than nucleotides. For example, if the site in question is TTT, then the three possible mutations are TTT->TCT, TTT->TGT and TTT->TAT and the relative frequency of each SNP is $p_{TTT\text{-}C}$, $p_{TTT\text{-}G}$ and $p_{TTT\text{-}A}$ respectively. The likelihood of a tri-allelic SNP being observed at this site is then simply:

$$Tri_{TTT} = (p_{TTT\text{-}C} \times p_{TTT\text{-}A}) + (p_{TTT\text{-}C} \times p_{TTT\text{-}G}) + (p_{TTT\text{-}G} \times p_{TTT\text{-}A})$$

This probability is then multiplied by the total number of TTT sites and repeated in a similar fashion across all triplet types to give the total number of tri-allelic sites expected. The process was repeated using estimated mutation rates from each gene rather than across all sites in the dataset. This incorporates the effects of any regional variation in triplet mutation rates.

4.3.3 Cryptic Variation

In order to investigate the effects of cryptic variation in the mutation rate on the number of tri-allelic sites we used the method described in chapter 2, but we only considered non-CpG human tri-allelic SNPs against chimpanzee non-CpG bi-allelic SNPs. We did not correct for the effects of adjacent nucleotides on the mutation rate since there is not enough data to estimate this for tri-allelic sites and the effects are small for bi-allelic

sites. As such, the expected number of coincident SNPs is simply the total number of alignments divided by the number of positions in the alignment that were not part of a CpG dinucleotide.

4.3.4 Origin of tri-allelic SNPs

To test whether tri-allelic SNPs in autosomal data could have been produced by a simultaneous mutation event or by an event linked to recombination we considered whether the minor alleles were significantly closer together on a phylogenetic tree of human haplotypes than would be expected by chance. For each tri-allelic site we took the 100 bi-allelic SNPs either side from each of the individuals sampled in the Environmental Genome Project and SeattleSNPs studies, not including the tri-allelic SNP itself. Where there were not 100 SNPs either side of the tri-allelic site within each gene, we used extended data either side of the tri-allelic site up to a total of 200 SNPs where possible. We then used PHASE (Stephens, Smith, and Donnelly 2001) to construct haplotypes from the variation data. The bi-allelic sites were bootstrapped 1000 times and each bootstrap dataset used to build a phylogenetic tree using the Neighbour-Joining method in Phylip (Felsenstein 2005). The tri-allelic site was then placed back on this tree. The distances between the minor alleles in the tri-allelic site were found by summing the lengths of branches on each tree separating the terminal nodes containing the alleles; if either one of the minor alleles was not a singleton, the distances between every pair of haplotypes containing the minor alleles was averaged. The expected distance between minor alleles was estimated by randomly placing two mutation events on two branches of each tree inferred from the bootstrapped data according to the inferred length of the branches; simulations in which the two mutations

fell on the same branch, or on the two branches descending from the root, were discarded since they would not generate a tri-allelic SNP. Minor alleles were designated as those at the lowest frequency and the average distance between them was calculated as before. The process was repeated across the 1000 bootstrapped trees for each tri-allelic SNP, and an estimated *p*-value was calculated as the proportion of trees in which the observed distance between minor alleles was smaller than the distance between the minor alleles of the simulated data. Fisher's combined probability test was used to calculate whether the *p*-values across all tri-allelic sites were significant.

In order to test whether the randomization procedure and analysis of phylogenetic trees was satisfactory we also derived the expected distance between a random pair of non-adjacent bi-allelic SNPs from within the set of haplotypes generated for each tri-allelic SNP, and then calculated whether these mutations fell significantly closer on the phylogenetic tree of haplotypes than we would expect by chance. In each case, phylogenetic trees were reconstructed as before, excluding the two randomly chosen bi-allelic SNPs. Where a single haplotype contained both minor alleles for the two SNPs chosen, the allele at lower frequency was used, thus generating three different alleles across all haplotypes for comparison. We found no significant difference between the real data and the simulated data ($p=0.28$) for the distances between minor alleles at the bi-allelic sites. We therefore conclude that our analysis procedure for phylogenetic trees is satisfactory and does not lead to artificial clustering of SNPs.

A second test was performed to judge whether the minor alleles of tri-allelic sites tended to cluster on a phylogenetic tree of haplotypes in the population. The distances between minor alleles of tri-allelic sites were calculated as above, however on this

occasion they were compared to the distances between minor alleles of tri-allelic sites that were generated by coalescent simulations. For each tri-allelic site the recombination rate was calculated using the Pairwise program in LDhat (McVean, Awadalla, and Fearnhead 2002), considering all haplotypes in the population and assuming a constant rate of recombination. A coalescent simulation was then performed using MS (Hudson 2002), which incorporated a model of demographic history as outlined by Adams and Hudson (2004) and the recombination rate and population structure for each particular tri-allelic site. Individuals were considered to be either African or non-African in the simulation. Finally, the program Seq-Gen (Rambaut and Grassly 1997) was used to generate haplotypes for each population under a finite sites model, with the mutation rate set such that the average nucleotide diversity would be 0.0015 (-s option). This is slightly higher than the average nucleotide diversity in humans, but was increased to reduce computing time. The process was repeated for each tri-allelic site until 100 data sets had been generated that contained a tri-allelic site; for each of these simulated tri-allelic sites we extracted the same number of bi-allelic SNPs as were present in the original data. Each set of sequences was then used as above, with the tri-allelic site removed, to generate a phylogenetic tree; the tri-allelic site was then placed back on the tree and the distance between minor alleles computed as above. The distribution of distances were then compared to the distribution of distances from the bootstrapped trees of the original data; the p-value was calculated by randomly pairing a value from the original boostrapped data, with a value from the coalescent simulations; the p-value was the number of times the former was less than the latter. The coalescent simulations depend upon the demographic model, but ethnic information for the DNA samples was only available for 60 of the 113 tri-allelic sites; we therefore only considered these.

In order to test whether tri-allelic SNPs are linked to recombination we calculated the average recombination rate across each gene in our data set using data from Kong *et al*. (2002). We split the genes into quartiles based on average recombination rate and tested whether the density of tri-allelic sites was significantly different between the upper and lower quartiles using a *z* test.

The expected number of tri-allelic SNPs that fall within immediately adjacent SNPs was calculated by multiplying the frequency of tri-allelic SNP sites by the frequency of immediately adjacent SNP sites (2 per pair of SNPs) within each intron.

4.3.5 Quantification

We estimate the relative contributions of single and simultaneous mutation events to the production of variation as follows. We assume the mutations are neutral, the population is stationary in size and the organism being considered is diploid. First, let us consider single bi-allelic SNPs. The expected number of bi-allelic sites in a sample of *n* sequences is:

$$S_s = 2N_e\mu_s \sum_{t=0}^{\infty} P(t,n) + 2N_e\mu_d \,.2\sum_{t=0}^{\infty} P(t,n)\,.(1-P(t,n)) \qquad (4.2)$$

where μ_s is the rate of single mutations, μ_d is the rate for simultaneous double mutations during the mitotic phase of germ line development, and $P(t,n)$ is the probability of observing a mutation that was produced t generations in the past in a sample of n sequences. Note that only simultaneous mutation events during mitosis are likely to generate two mutations that can both be inherited; this is because only one meiotic product generates an egg in females so only one mutation from a simultaneous event

during meiosis will be inherited. Furthermore, human females typically only have one offspring at a time; hence, only one product from a simultaneous event in male meiosis will be transmitted. The first summation denotes the probability of observing a SNP produced by a normal mutational event, and the second summation denotes the probability of observing a bi-allelic SNP that was originally produced by a simultaneous event, with one allele being lost through genetic drift and therefore only contributing a bi-allelic SNP to the population.

The expected number of tri-allelic SNPs is approximately:

$$S_t = 2N_e\mu_d \sum_{t=0}^{\infty} P(t,n)^2 \qquad (4.3)$$

This is only an approximation because it assumes that the frequencies of the two mutations are independent, whereas they are not; for example, if one allele goes to fixation then the other allele can no longer exist. However, this approximation is likely to be good since the new mutations will generally be rare.

The probability of observing a SNP in the population, $P(t,n)$, can be split into two components; the probability that a SNP is segregating in the population, $y(j,t,N)$, where j is the number of copies of the new allele in the population of size N, and the probability that it is sampled in our data, $z(n)$. We can estimate $y(j,t,N)$ using a transition matrix approach as follows. We initially introduce a single mutation into our population: $y(1,0,N) = 1$ and $y(j,0,N) = 0$ for j>1. The probability of the mutation being at a frequency j given that we had i copies in the previous generation can be calculated from the binomial distribution:

$$X(i,j,N) = \frac{n!}{j!(n-j)!}\left(\frac{i}{2N}\right)^j\left(1-\frac{2N-i}{2N}\right)^{2N-j}$$

so

$$y(j,t+1,N) = \sum_{i=1}^{2N-1} y(i,t)X(i,j,N)$$

The chance of sampling a SNP is:

$$z(n) = 1 - (\frac{j}{2N})^n - (1 - \frac{j}{2N})^n$$

where n is our sample size, $(j/2N)^n$ is the chance that one of the minor alleles gets sampled in all cases, and $(1-j/2N)^n$ is the chance that the other allele gets sampled in all cases. The likelihood of observing the SNP is therefore:

$$P(t,n) = \sum_{j=1}^{2N-1} y(j,t,N).z(n)$$

Both equations 4.2 and 4.3 involve infinite sums; to determine a reasonable limit of this summation we note that $\sum P(t,n)$ should be equal to $\sum(2/i)$ as given by Watterson's classic formula for the number of neutral polymorphisms segregating in a sample of sequences (Watterson 1975):

$$S_w = 4N\mu \sum_{i=1}^{2N-1} \frac{1}{i}$$

The convergence of $\sum P(t,n)$ depends upon the number of chromosomes sampled; we required that $\sum P(t,n)$ was within 1% of $\sum(2/i)$.

From equations 4.2 and 4.3 it is straightforward to estimate the relative rates of single and simultaneous mutation, μ_s and μ_d, from the observed numbers of bi-allelic and tri-allelic sites, S_s / S_t.

4.4 Results

4.4.1 Excess of tri-allelic sites

We used data from 896 nuclear genes that had been resequenced in between 90 and 95 human individuals to search for tri-allelic sites. After removing CpG and coding sites we had a total of 36,702 transitions, 20,375 transversions and 113 sites that had three alleles segregating in the human population (appendix 4.1). This is significantly greater than the 61.15 tri-allelic sites expected by chance if mutations are randomly distributed across non-CpG sites (ratio of observed over expected = 1.85, with a standard error of 0.17, $p < 0.001$ under the null hypothesis that the ratio is one). We also searched for tri-allelic SNPs at four fold synonymous sites in human mitochondrial genes in 4764 complete sequences. We found 1125 transitions, 173 transversions and 126 tri-allelic sites. In this case we found no significant excess of tri-allelic sites above that expected by chance (observed over expected = 1.20, $p > 0.05$). We do not consider the results from mtDNA further. The excess of tri-allelic SNPs in nuclear DNA could be potentially caused by one of four processes; sequencing errors, natural selection, increased rates of mutation at single nucleotides or another form of mutation that may include mismatches at recombination or a primary mutation inducing a second mutation on the strand opposite.

4.4.2 Sequencing Error

It is possible that some tri-allelic sites are the result of sequencing error and this may lead us to over-estimate the excess of tri-allelic SNPs, however this seems unlikely for

the following reasons. There are only two ways in which sequencing error could generate the excess; first, if polymorphic sites tend to be more prone to sequencing error, and second if some sites tend to be prone to sequencing error and two of the three alleles at a tri-allelic site are generated by error. However, there is no evidence and no reason to believe that sites that are polymorphic tend to be more prone to sequencing error, and very few of our tri-allelic sites have two singleton mutations, as we would expect if two alleles at tri-allelic sites were generated by error. However, to investigate the matter further we performed an additional analysis. Should the excess of tri-allelic sites be due to sequencing error we would expect to see a higher frequency of singletons at tri-allelic sites than we do at bi-allelic sites in the observed data. To test this, we selected one of the minor alleles at each tri-allelic site at random, since bi-allelic sites only contain two alleles, and then compared the frequency of singletons at randomly chosen alleles at tri-allelic sites with the frequency of singletons found at bi-allelic sites; we bootstrapped the data 1000 times to obtain an average singleton frequency and confidence intervals at tri-allelic sites. The frequency of singletons at bi-allelic sites is 0.347 (95% confidence interval: (0.342,0.352)) and the average frequency of singletons at tri-allelic sites is 0.345 (0.257,0.434), which are not significantly different ($p=0.44$). Furthermore, we have contacted the researchers who produced the data and they confirm that the validation process is more rigorous for tri-allelic sites due to their rarity and interesting nature. Tri-allelic sites are confirmed on both strands of the DNA duplex and additional PCR reactions are performed to show that the tri-allelic site is present in at least two different amplicons (Rieder 2010). Thus, there is no evidence that the excess of tri-allelic sites is caused by sequencing error.

4.4.3 Natural selection

Selection is expected to lead to an apparent excess of tri-allelic sites because SNPs will not tend to segregate within regions under selection, and therefore SNPs will appear to be clustered between these areas. Firstly, this is unlikely to be the case here as all of the sequences considered are intronic, and although selection is known to act in these regions, it is thought to only affect a small percentage of sites (Waterston et al. 2002; Dermitzakis, Reymond, and Antonarakis 2005; Asthana et al. 2007). Furthermore, if selection were responsible for the excess of tri-allelic SNPs we would expect to see SNPs clustering more generally. However, if we look at the distances between SNPs, and compare this to the results from simulations in which SNPs are randomly distributed, then we see no evidence of clustering except an excess of tri-allelic sites and an excess of immediately adjacent SNPs (Figure 4.2). We consider the excess of adjacent SNPs separately below. The distances between SNPs suggest that selection is not impacting on the number of tri-allelic sites present.

4.4.4 Mutation Hotspots

The excess of tri-allelic SNPs could be a result of local variation in the mutation rate in the human genome. It has previously been shown that the mutation rate varies as a function of local context effects, particularly depending upon the adjacent nucleotides (Blake, Hess, and Nicholsontuell 1992; Zhao et al. 2003; Hwang and Green 2004). Such variation in the mutation rate could lead to an increased number of tri-allelic SNPs if some sites have an elevated mutation in two or more pathways; e.g. if both C->T and C->A occur at higher rates. To investigate whether neighbouring nucleotide effects

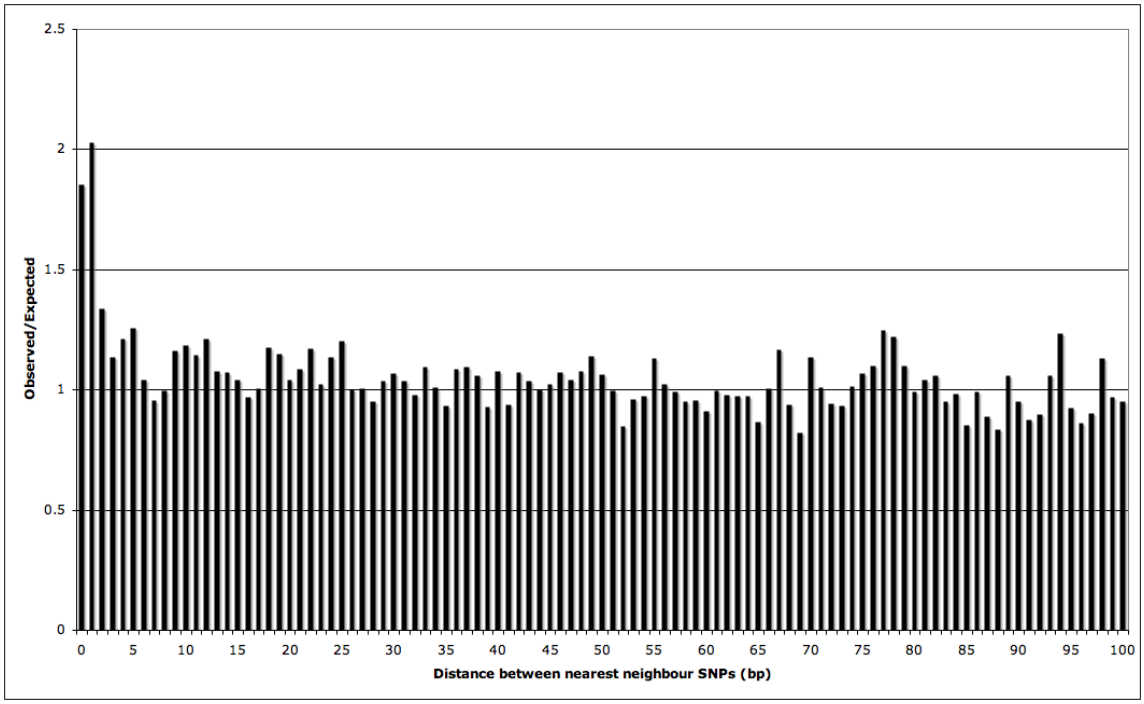


Figure 4.2. Observed over expected values for the distance to the nearest neighbour SNP within each intron.

could cause the excess of tri-allelic sites, we estimated the probability of observing a SNP at the central nucleotide of each triplet, ignoring CpGs, and used these to estimate the expected number of tri-allelic sites. If we estimate the probabilities across all our genes we infer the expected number of tri-allelic sites to be 61.88; this is only slightly larger than the estimate ignoring adjacent nucleotide effects, and is significantly less than the number observed (observed/expected = 1.83, standard error = 0.17, $p < 0.001$). If we estimate probabilities within genes, thus controlling for any regional variation in mutation rates within chromosomes, this expectation increases slightly to 69.03, but this is still highly significantly different from the observed number ($p < 0.001$). Local context effects are therefore not the cause of the excess of tri-allelic sites. We do not consider the effects of nucleotides further away as these have been shown to have a much smaller effect on mutation rates than adjacent nucleotides (Krawczak, Ball, and Cooper 1998; Zhao et al. 2003), which themselves have little impact on the expected number of tri-allelic SNPs.

It is also possible that the excess of tri-allelic SNPs is caused by CpG alleles that were subsequently lost from the population. If we consider a CpG site that mutates at a high rate to produce a TpG and an ApG; if the CpG is then lost from the population and either the TpG or ApG then mutates to a GpG, this generates a tri-allelic site which was in part caused by the increased mutation rate associated with a CpG. In order to test for this we repeated the analysis and excluded all sites preceded by a C or followed by a G; we find that the effect is still significant (observed/expected = 1.77, standard error = 0.18, $p<0.001$). Therefore, CpGs that have been lost are not causing the excess of tri-allelic sites.

However, we have previously shown that the mutation rate varies across the human genome in a cryptic nature that is not associated with any specific context effect (chapters 2 and 3). This was demonstrated by showing that there is an excess of sites with a SNP at the orthologous position in humans and chimpanzees (coincident SNPs). Such cryptic variation could potentially lead to an excess of tri-allelic sites, should at least two mutational pathways be elevated at particular sites (e.g. a transition and a transversion). However, we note that this cryptic variation seems to largely elevate the rate of mutation of a single mutational pathway; there is a large excess of cases in which an X<->Y SNP in humans is coincident with an X<->Y SNP in chimps, but little or no excess of X<->Y in humans and X<->Z in chimpanzees (where X,Y and Z can be A,T,G or C). Cryptic variation would therefore not tend to generate tri-allelic sites. However, to investigate the matter further we considered whether tri-allelic sites in humans also tended to have a SNP at the orthologous position in chimpanzee, and calculated the number expected assuming that chimpanzee SNPs and tri-allelic sites were randomly distributed relative to one another. As an excess of tri-allelic SNPs requires an increased level of mutation along two pathways, we would expect the ratio of coincident tri-allelic SNPs to their expected number to be far in excess of that found for coincident single SNPs. Because we are not interested in the excess of tri-allelic sites *per se* we can take advantage of tri-allelic sites found in dbSNP; there are ~50,000 examples. These were BLASTed against a dataset of chimpanzee SNPs to yield a data set of 548 alignments of 81bp with the chimpanzee SNP in the central position and the human tri-alllelic SNP elsewhere within the alignment. Of these alignments, 17 have the human and chimpanzee SNPs in the same position, as opposed to the 6.96 we would expect were the SNPs distributed at random (observed over expected ratio = 2.44); this is not significantly different from the ratio for non-CpG single SNPs in which local

context effects are ignored, which is 2.40 (chapter 2). There is therefore no evidence that cryptic variation in the mutation rate tends to generate an excess of tri-allelic sites.

There are a number of additional mutation hotspot mechanisms that could cause an excess of tri-allelic sites, which need to be considered. First, it has been suggested that sequences adjacent to indel events may have elevated rates of mutation; this is most evident up to 100bp from the indel, but effects decline away from indels across several hundred base pairs (Tian et al. 2008). It is unlikely that an increase in the rate of mutation in these areas could cause the excess of tri-allelic SNPs in our data because if the effect were sufficiently large, we would expect to see a clustering of SNPs in certain parts of the genome. It is clear from figure 4.1 (discussed in the section on natural selection) that this is not the case. Furthermore, alpha-polymerase pause sites are also thought to mutate at a higher than normal rate (Todorova and Danieli 1997), and could cause an excess of tri-allelic sites. In our data however, there is only one occurrence of the motif associated with a pause site in the region immediately upstream of tri-allelic sites and thus the motif does not impact on our results.

4.4.5 Recombination and simultaneous mutation

The evidence above suggests that it is not particular sites that tend to produce tri-allelic SNPs, so maybe tri-allelic sites are generated by a mechanism which can occur at all sites with a similar probability, but one in which one mutation generates the second mutation. There are at least two related potential mechanisms. First, it is possible that a mutation could induce a subsequent mutation, possibly through the formation of heteroduplex DNA during recombination. The lack of an excess of tri-allelic sites in

mitochondrial DNA, in which recombination is rare or absent, would be consistent with this model. Second, it might be possible that two mutations occur within a single DNA duplex; for example, a G=C base pair might become an A=C mismatch, which then becomes an A=G. If replication runs through this mismatch before it is repaired we would end up with the G=C allele being mutated to A=T and C=G; so two new mutations have been generated within a single event. Alternatively, it may be possible that both strands of DNA are mutated simultaneously, perhaps as a result of a chemical or radiation event. This process of simultaneous mutation would need to occur in the mitotic phase of germ line development in order for the two new alleles to be potentially transmitted to the next generation. We refer to these as the "recombination" and "simultaneous mutation" models respectively.

In order to investigate whether one of these two mechanisms generates the excess of tri-allelic sites we can potentially test a prediction that both models make for recombining data; they both predict that new mutations should cluster together on a phylogenetic tree of human haplotypes. We expect this clustering under the simultaneous mutation model because we hypothesise that both new mutations will be produced on the same genetic background. The clustering is also expected under the recombination model because the first mutation should induce the second mutation equally as often on the original haplotype as it does on another haplotype in the population. To investigate whether we could detect clustering in recombining autosomal loci we took each of our tri-allelic sites and 200 bi-allelic SNPs surrounding the site where possible; using the bi-allelic SNPs we constructed haplotypes and inferred the phylogenetic relationships between them. We then calculated the average distance between haplotypes containing the minor alleles. Amongst our 113 tri-allelic sites we find 6 cases in which the minor

alleles are significantly closer to each other at the 5% level than if mutations are placed on the phylogenetic trees at random, approximately what we would expect by chance alone; however if we combine probabilities across all tri-allelic sites we find significant evidence for proximity of the minor alleles ($p<0.05$). Nevertheless, comparing the observed distances between minor alleles with those generated by randomly placing mutations on the same phylogenetic tree may not be the most appropriate way to detect clustering. There may be some genealogies that tend to produce tri-allelic sites by double mutation, where the minor alleles tend to be clustered together on deeper branches of the phylogenetic tree. As a consequence, the method above may average across several genealogies if there has been recombination and so produce an average genealogy that does not tend to lead to an excess of tri-allelic sites. Thus we performed coalescent simulations under a realistic demographic model, with the rate of recombination estimated from the bi-allelic sites, to generate a set of simulated tri-allelic sites for each observed tri-allelic site. We then compared the observed distances between the minor alleles of tri-allelic sites with those produced from the coalescent simulations. We only knew the ethnicities of the individuals sequenced for 60 of the tri-allelic sites and so could only perform simulations for those sites because of the need to incorporate demography; in 12 cases the minor alleles are significantly closer to each other at the 5% level than those generated from coalescent simulations. If we combine probabilities across all 60 tri-allelic sites we find highly significant evidence of proximity ($p<0.001$). This is consistent with both the simultaneous mutation and recombination pathways.

In principle it is possible to differentiate between the recombination and simultaneous models by considering non-recombining nuclear DNA, such as the non-recombining portion of the Y-chromosome (NRY). In the NRY, under the simultaneous mutation hypothesis we expect both mutations to appear at the same time on the same genetic background in about half the tri-allelic sites, the other half being a consequence of chance alone. The simultaneous generation of two mutations will manifest itself as two new alleles emanating from a single node in the phylogenetic tree of human haplotypes. Unfortunately, to our knowledge only a single non-CpG tri-allelic SNP has been reported for the NRY; this is an A,T,C tri-allelic SNP termed M116 (Underhill *et al.* 2001). In this case, the two minor alleles, T and C, are found in different haplogroups and as such cannot have been caused by a simultaneous mutation event. This does not disprove that tri-allelic sites are produced by simultaneous mutation since we infer that about 50% of tri-allelic sites are due to chance alone (see above).

Alternatively, we may be able to differentiate between the two models by considering the prediction that under the recombination model, tri-allelic SNPs and recombination rates should be correlated. We have already shown that there is no excess of tri-allelic sites in human mitochondrial DNA and clearly there is a prediction under the recombination model that there should be no excess in non-recombining sequences. However, in mtDNA a lack of tri-allelic sites does not necessarily point to tri-allelic SNPs being generated during recombination, as there are many other factors that differentiate the mutation process in mtDNA and nDNA that could be equally likely to generate the result. Consequently, we tested for a correlation between tri-allelic SNPs and recombination rates in the autosomal datasets. We separated our data set of introns from genes into quartiles based on the average recombination rate across the gene (rates

taken from Kong *et al.* (2002)) and found that there was no significant difference between the number of tri-allelic SNPs per sampled site in genes that were in the upper and lower 25% of recombination rates ($p = 0.77$). We also tested for a correlation between recombination rate and the presence/absence of a tri-allelic site across genes using logistic regression; there was no evidence of a significant correlation ($p=0.99$). It should be noted at this point that our test for a correlation between tri-allelics SNPs and recombination may not include all gene conversion events, as the genetic map only measures crossover rates and our hypothetical recombination mechanism would also apply to gene conversion events. However, as gene conversion and crossover hotspots tend to coincide (Jeffreys and May 2004), it is likely that the result would be mirrored when considering gene conversion as an indicator of tri-allelic SNP density. There is therefore no evidence that tri-allelic sites are linked to recombination. This leads us to believe that the simultaneous mutation model most likely explains the excess of tri-allelic SNPs in the human genome.

4.4.6 Adjacent mutations

As we noted above, besides an excess of tri-allelic sites there is also an excess of adjacent SNPs. It has been previously noted that adjacent substitutions are more common than you would expect by chance (Averof *et al.* 2000) and it has been suggested that this is due to the simultaneous mutation of adjacent nucleotides. To investigate whether this is the case we compared the absolute difference in minor allele frequency (MAF) between adjacent SNPs. If adjacent SNPs are produced simultaneously then we expect many adjacent SNPs to have identical MAF since they can only differ in frequency if they are broken up through recombination. This is what

we observe; approximately half of all adjacent SNPs have identical MAFs (252/506), which is consistent with the observation that adjacent SNPs are approximately twice as common as expected by chance. We also note that the absolute difference in MAF between adjacent SNPs is significantly smaller than the absolute difference in MAF between one of the adjacent SNPs (randomly chosen) and the next non-adjacent SNP ($p<0.001$, average difference in MAF for adjacent SNPs = 0.073, average difference in MAF for non-adjacent SNP = 0.107). Thus, it seems that there is a process that produces adjacent SNPs simultaneously, and it therefore seems possible that a similar process could also generate tri-allelic sites if a mutation event that causes a doublet mutation along a strand can also cause a double mutation across strands, which could occasionally occur at the same time. To investigate this we searched our data for any case in which a tri-allelic site was adjacent to another SNP. We found one case, and although this feature is rare, this is significantly higher than the 0.008 we expect by chance alone ($p < 0.01$). The coincidence of a tri-allelic site and an adjacent SNP could be due to either some sites having a greater chance of producing adjacent and tri-allelic sites, or due to the generation of both simultaneously.

4.5 Discussion

We have shown that there is an excess of sites in the human genome that have three alleles segregating in the population. The excess cannot be explained by sequencing errors, natural selection or an increased mutation rate for single events at particular sites. Instead, there is some evidence that a proportion of tri-allelic sites may be caused by a single mutation mechanism in which two new alleles are produced at the same or

similar time, on the same or similar genetic background; the minor alleles at a tri-allelic SNP tend to be closer together on the phylogenetic tree than one would expect by chance. We show that the clustering is unlikely to be caused by a mutational mechanism linked to recombination, as there is no association between recombination rates and genes that contain tri-allelic SNPs. We have also shown that there may be an association between tri-allelic and immediately adjacent SNPs. None of these lines of evidence is individually particularly strong, but collectively they suggest that a proportion of tri-allelic sites are a consequence of simultaneous mutation. A conclusive test can be made using Y-chromosome data, and the 1000 human genome project is likely to provide sufficient information to resolve the problem since the project will produce long non-recombining Y-chromosome sequences from many males. However, this data is unlikely to be available for another 12 to 18 months (McVean 2009).

Although we do not know the specific mechanism that generates two new alleles at a single site, there are a number of potential candidates. First, it is possible that double strand breaks that occur as a result of both endogenous and exongenous factors, could result in nucleotides at the same site on opposite strands being more prone to mutation. It seems unlikely that this mechanism is the one responsible here, since double strand breaks are generally considered to be mutagenic over a larger distance than at single sites (Pfeiffer 1998), and we do not observed clustering of mutations in the regions surrounding tri-allelic sites. Second, it is possible that error-prone polymerases could increase the likelihood of a second mutation opposite a site that has already undergone a mutation. It is known that Y-family polymerases have a high error rate when replicating undamaged DNA (Goodman 2002; Rattray and Strathern 2003) and furthermore, particular polymerases are known to preferentially insert non Watson-

Crick base pairs at lesions (Goodman 2002; Rattray and Strathern 2003). For example, polymerase iota preferentially inserts a guanine opposite a thymine when recruited to a lesion site (Johnson et al. 2000); should the original mutation have been A->G, this kind of error would lead to an A, G and C being present at the same site across different copies of the DNA after replication. Whether this specific polymerase is involved at tri-allelic sites is not known, however it is not difficult to imagine that a similar mechanism involving other enzymes could play a role. A similar mechanism has been invoked to explain why rates of CpG and non-CpG substitutions are correlated across genes in the human genome (Walser, Ponger, and Furano 2008). The idea is further supported as translesion polymerases are known to be highly expressed in germ cells (Rattray and Strathern 2003), where the generation of a tri-allelic site would need to occur in order to be inherited and become visible in the population. Third, it may be possible that tri-allelic sites are generated at dual abasic sites, where the nucleotides on both strands of the duplex become lost. Were replication to encounter such a problem, it could lead to two different mutations at the same site. It has been proposed that the *in vivo* steady-state level of abasic sites is between 4,500 and 200,000 for a 3 billion nucleotide genome (Nakamura and Swenberg 1999; Atamna, Cheung, and Ames 2000) and even if the actual number is at the lower end of these estimates, the rate of depurination could be increased by chemical or enzymatic induction, such as free radicals or alkylating agents; the latter of which has been shown to accelerate the rate of nucleotide loss by 6-fold (Wilson and Barsky 2001). Under normal conditions, the abasic site would be incised by AP endonucleases and repaired using the nucleotide opposite as a template, however should both nucleotides be lost there would be no template to re-generate the original nucleotide. It has also been shown that endonucleases do not need the base opposite for recognition of an abasic site (Wilson et al. 1995). Fourth, it could

potentially be the case that two mutations on opposite strands are caused by chemical or radiation event attacking both strands simultaneously, however in the case of the latter it is difficult to see how this could occur in germ line cells that are generally less prone to attack by external agents. Finally, it could be the case that a primary mutation induces a mutation at the site opposite if the bond between the two nucleotides is lost and the site essentially becomes single stranded DNA (ssDNA). Cytosine bases have been shown to be more prone to deamination and subsequent mutation on ssDNA when adjacent to a guanine residue (Frederico, Kunkel, and Shaw 1990) and although we have excluded CpG sites in this analysis, a similar process is thought to occur during somatic hypermutation where an associated enzyme, activation induced deaminase (AID), induces deamination at cytosines in different contexts on ssDNA (Rattray and Strathern 2003). However, although a similar mechanism could occur here, we would likely observe a general clustering of mutations in the region, which we do not see.

The available evidence suggests that the excess of tri-allelic sites is caused by the simultaneous production of two new alleles in a single individual. If this is the case then we can estimate the frequency of this process using equations 4.2 and 4.3. To a first approximation the relative numbers of bi-allelic and tri-allelic SNPs depend upon their mutation rates and the relative probabilities of detecting a single SNP ($\Sigma P(t,n)$), and two SNPs generated simultaneously ($\Sigma P(t,n)^2$). Surprisingly we find that $\sum P(t,n)^2$ is only about fifteen-fold lower than $\sum P(t,n)$ when large numbers of chromosomes have been sampled (table 4.1); so the chance of sampling two mutations produced in the same generation is actually quite high in the data we have analysed.

Number of Chromosomes	$\sum 2/i$	$\sum P(t)$	$\sum P(t)^2$	$\sum P(t) / \sum P(t)^2$
4	3.66	3.62	0.017	212.02
10	5.65	5.58	0.049	112.70
40	8.51	8.29	0.190	43.73
100	10.35	9.88	0.419	23.61
200	11.75	10.90	0.698	15.61

Table 4.1. The summed probabilities that a single SNP is sampled in *n* chromosomes until it is fixed or lost, and the summed probabilities that two mutations produced simultaneously are sampled, together with a ratio of the two values across different numbers of sampled chromosomes.

We can use equations 4.2 and 4.3 to estimate the ratio of the rates of single and simultaneous mutation, μ_s/μ_d as follows. We infer that approximately half of all tri-allelic sites are a consequence of simultaneous mutation, hence S_d = 51.13 and S_s = 57077. In our data between 180 and 190 chromosomes have been sampled so $\sum P(t)$ / $\sum P(t)^2$ is between 16.05 and 16.54 and hence μ_s/μ_d is between 65.6 and 67.7; so single mutations occur approximately 65 times more frequently than simultaneous mutations and since each simultaneous event produces two new mutations, we estimate that about 3% of all distinct SNPs are generated in this fashion.

We have also shown that there is an excess of adjacent SNPs, and that at least half of these adjacent SNPs appear to be generated simultaneously. If we assume that the doublet mutations remained linked throughout their life – i.e. there is no recombination between them, then we can directly estimate the rate of adjacent mutation (μ_a) by considering the ratio of single SNPs to adjacent SNPs; using this approach we estimate μ_s/μ_a as 225.6, assuming that approximately half of all immediately adjacent mutations occur simultaneously. As adjacent mutation events contribute two new SNPs, we estimate that about 0.89% of all distinct SNPs are generated in this fashion. So although adjacent SNPs are slightly more common than tri-allelic sites, the rate at which they are produced is actually lower, and this is because the probability that two independent SNPs survive to be sampled is considerably lower than the probability that two linked SNPs survive (table 4.1). Of course, this figure depends on the rate of recombination between adjacent SNPs; should this rate be extremely high, there will be almost no linkage between adjacent SNPs, if it is low, adjacent SNPs will behave in the manner of single SNPs. In this case it is reasonable to suggest that the latter is most probably more realistic. Our estimate of the ratio of single over doublet mutation rates

of 225.6 is closer to the ~1000 estimated by Kondrashov (2003) than the ~10 estimated by Averoff *et al.* (2000). We believe that our estimate is likely to be most accurate as it uses the most direct approach to compare mutation rates in neutral sequences.

Acknowledgements. We are very grateful to Phil Green, Mike Zody and an anonymous referee for comments. AH and AEW were funded by the BBSRC and AEW by the European Community and the National Evolutionary Synthesis Center.

5. The local context and genomic distribution of cancer mutations

5.1 Abstract

Recently the genome sequences from four cancer cells have been published, along with the genome from a non-cancer tissue from the same individual, allowing the identification of new somatic mutations in the cancer. The data comes from two acute myeloid leukemia (AML) cells, a small-cell lung cancer (SCLC) cell and a skin cancer cell. Here we investigate the patterns of mutation in the cancer tissues as well as the distribution of those mutations at a number of different scales; we compare these to the patterns seen in the germ-line. The pattern of mutation in SCLC and skin cancer is dominated by mutations associated with specific mutagens, as others have shown before. Interestingly one of the two AML cancers shows a very similar pattern to the SCLC, suggesting that it might have been caused by similar mutagens. In contrast the second AML cancer shows a pattern that is very similar to that observed in the germ-line. The rate of mutation at individual sites is affected by the identity of the adjacent nucleotides in all cancers, and these patterns are very similar to those observed in the germ-line, except in the skin cancer genome, where patterns of mutation are very different to the germ-line. All cancer genomes show substantial variation in the number of mutations per MB, and these patterns are quite strongly correlated between SCLC and skin cancers, and between the two AML genomes; all cancers are significantly

correlated to the germ-line. We find no evidence that genes associated with cancers are found disproportionately in regions with high mutation rates. Finally we show that the mutation rate differs between chromosomes for all cancers; echoing the results at a MB scale, mutation rates were most strongly correlated between skin and SCLC, and between AML genomes, and also the germ-line.

5.2 Introduction

Mutations can be divided into those that occur in the germ-line and those that occur in the soma. Germ-line mutations contribute to future generations and are the basis of inherited genetic disease as well as being the raw material of evolutionary change. In contrast, somatic mutations do not contribute to future generations but can give rise to cancers and most probably contribute to ageing (Vijg 2000; Finkel, Serrano, and Blasco 2007). It has been estimated that on average each of us receives approximately 50-100 new germ-line mutations from our parents (Kondrashov 1995; Lynch 2010), but this is completely dwarfed by the number of somatic mutations a cell will accumulate during its life-time within an individual; it is thought that cells in proliferating tissue are likely to have accumulated 10,000s of new somatic mutations by the time an individual reaches mid-life (Lynch 2010). Cancers are no different to other somatic tissue and accumulate thousands of mutations by the time a diagnosis is made. Recently the sequence from four cancer genomes have been published, along with the genome from non-cancer tissue from the same individual, allowing the identification of new somatic mutations in the cancer. The data comes from two acute myeloid leukemia (AML) genomes (Ley et al. 2008; Mardis et al. 2009), henceforth referred to as AML1 and

AML2 respectively, a small-cell lung cancer (SCLC) genome (Pleasance et al. 2009b) and a skin cancer genome (malignant melanoma) (Pleasance et al. 2009a).

In the germ-line transitions occur approximately 2 times as frequently as transversions, and GC->AT changes occur much more often than the reverse (Gojobori, Li, and Graur 1982; Li, Wu, and Luo 1984; Blake, Hess, and Nicholsontuell 1992). In contrast, it has been shown that lung and skin cancer genomes show evidence of the mutagenic mechanisms that are thought to cause these diseases; the lung cancer genome shows a large excess of GC->TA transversions (Pleasance et al. 2009b), which is thought to be associated with tobacco smoke carcinogens (Pfeifer et al. 2002), and the skin cancer shows a very large excess of CG->TA transitions (Pleasance et al. 2009a), which are thought to be a consequence of UV light (Pfeifer, You, and Besaratinia 2005). The pattern of mutation has not been studied in the AML genomes.

The rate of mutation at a site is also known to depend on the identity of the neighbouring nucleotides in the germ-line, the most dramatic example being CpG dinucleotides, which are 10-15 times more mutable than other nucleotides (Coulondre et al. 1978; Bird 1980). However, the mutation rate varies by at least 2-3 fold at all other sites as a function of the adjacent nucleotides (Hess, Blake, and Blake 1994; Zhao and Boerwinkle 2002; Hwang and Green 2004). Strong context effects have also been noted in the skin cancer genomes; for example, 92% of C->T transitions in skin cancer occur at sites proceeded by another pyrimdine, compared to the ~50% expected by chance (Pleasance et al. 2009a). This pattern is consistent with the action of ultra-violet light in generating pyrimidine dimers, and also the impacts of reactive oxygen species that preferentially target CCN and NCA triplets, where the central base is mutated and

N is any nucleotide (Pfeifer, You, and Besaratinia 2005). Again, no neighbouring nucleotide analysis has been performed on the mutations in AML genomes.

Finally, the mutation rate has been shown to vary in the germ-line at both a chromosomal (Lercher, Williams, and Hurst 2001; Li, Yi, and Makova 2002; Gaffney and Keightley 2005) and a sub-chromosomal scale (Wolfe, Sharp, and Li 1989; Wolfe and Sharp 1993; Matassi, Sharp, and Gautier 1999; Gaffney and Keightley 2005). The reasons for this variation are poorly understood, however, the fact that some regions of the genome can have high mutation rates opens up the possibility that some genes tend to be involved in cancer simply because the genomic region in which they are located has a high mutation rate.

Here we perform a comparative analysis of the mutation patterns observed in the four cancer genomes and in the germ-line. The four genomes provide a unique opportunity to fully compare and contrast the mutation signatures associated with three different types of cancer, and also those present in two different individuals suffering from the same disease: AML. We consider the patterns of mutation at three scales; first at single nucleotides, incorporating the effects that adjacent and adjacent-but-one nucleotides have on the rate of mutation, and also at a 1MB scale and on a chromosomal level. We chose a scale of one megabase since Gaffney and Keightley (2005) have shown that this is the scale over which the germ-line rates vary on a sub-chromosomal level. Our overall aims are two-fold. First, we hope that studying the patterns of mutation in the germ-line and somatic tissue will allow a greater understanding of why the mutation rate varies across the genome. Second, we wish to test whether certain genes are involved in cancer because they have high rates of mutation.

5.3 Materials and Methods

5.3.1 Germ-line mutation rates

Germ-line mutation rates were calculated as follows. We downloaded human SNP data from the SeattleSNPs project (SeattleSNPs 2008) and Environmental Genome Project (NIEHS-SNPs 2008), and removed all coding sites. SNPs were orientated using orthologous chimpanzee sequences that were downloaded using Ensembl Biomart (www.ensembl.org/biomart/martview/) and aligned to human sequences using FSA (http://math.berkeley.edu/~rbradley/papers/manual.pdf), which incorporates Exonerate (Slater and Birney 2005) and MUMmer (Kurtz et al. 2004). We were unable to find a small number of orthologous chimpanzee sequences (~6% of human SNPs); the major allele was used to infer the ancestral state in these cases. To calculate single nucleotide mutation rates we tallied the number of each type of nucleotide across the human sequence data, ignoring any regions that were not scanned for variation, and then calculated the frequency of each type of SNP, μ_{X-Y} (where X and Y are either A, G, T or C), by dividing the number of SNPs where the ancestral allele was X by the total number of X nucleotides in the human sequence data. For example, μ_{A-G} was estimated by dividing the number of sites where the inferred mutation was from A to G (i.e. A was the allele present at the orthologous chimpanzee position and G was the second allele at the SNP site) by the total number of sites that were A. A similar approach was used to incorporate the effects of neighbouring nucleotides; triplet mutation rates, $\mu_{NXN-NYN}$ (where N is either A, G, T or C), were calculated by dividing the total number of SNPs that were NXN to NYN by the total number of triplets that were NXN in the human sequence data. Reverse complement triplets and mutations were considered to

be the same (e.g. TTT>C and AAA>G are considered to be equivalent and were summed). To isolate the independent effects of neighbouring nucleotides, triplet rates were divided by the overall rate for mutation of the central nucleotide, regardless of context – e.g. TTT>C was divided by the rate of T>C, which is the average rate of T>C mutations across triplets.

To estimate patterns of mutation in the germ-line on a 1 MB and chromosomal scale, we estimated the divergence between the human and chimpanzee genomes as follows. Alignments using the GRCh37 version of the human genome and PanTro2 version of the chimp genome were downloaded from the UCSC website (http://genome.ucsc.edu/) and nucleotides were masked where the quality score was less than 40 in the chimpanzee genome (representing an error rate of 1/10000). Quality scores were unavailable for chromosomes 21 and Y, and so they did not undergo quality score masking. In order to minimize the possibility of non-homologous sites contributing to divergence data, we masked any regions that contained more than 10% divergence across 100bp, with sliding windows every 10bp. We also repeated the analysis using a 5% divergence threshold and obtained very similar results. Finally, we removed any regions of less that 20bp that were flanked both sides by more than 40bp of gap sequence, as we could not be confident in the reliability of these regions. Substitution density per megabase (MB) was then calculated only in regions that contained at least 100kb of unmasked sequence. Any regions below this threshold were excluded from further analysis; these were typically heterochromatic regions near centromeres. Chromosome mutation rates were calculated by dividing the number of substitutions per chromosome by the total number of unmasked nucleotides.

5.3.2 Cancer mutations

SCLC and skin cancer mutations were downloaded from the supplementary sections of the respective papers (Pleasance et al. 2009a; Pleasance et al. 2009b). The two AML genomes were obtained via dbGaP (http://www.ncbi.nlm.nih.gov/gap/), through dbGaP accession number phs000159.v1.p1. For SCLC and skin cancer, the data included mutations that were found only in the cancer cells, with any positions that were already present in dbSNP also removed. To ensure consistency across datasets, we also took the same approach for the AML genomes, removing mutations present in normal cells, dbSNP and the Watson and Venter genomes. For AML1, the data remaining is the same as that published in the genome paper, however, for AML2 the original study also removed some SNPs found in the 1000 genomes project (http://www.1000genomes.org). These were not removed in our analysis, to maintain consistency across the cancer genomes, however the mutation patterns are very similar for AML2 on all scales both with and without these SNPs (correlation considering neighbouring nucleotide effects, $r = 0.995$, $p<0.01$; correlation on 1MB scale: $r = 0.947$, $p<0.01$; correlation on chromosome scale: $r = 0.992$, $p<0.01$). Mutation rates for the central nucleotide of a triplet were calculated by dividing the number of occurrences of each triplet-mutation type by the number of occurrences of the reference triplet in the NCBI36 genome sequence (downloaded from the UCSC website). For example, the mutation rate of TTT>C was calculated by tallying the number of times TTT>C was observed in the cancer genome, and dividing this by the number of times TTT occurred in the NCBI36 reference sequence. Again, to determine the independent effects of neighbouring nucleotides, triplet rates were divided by the overall rate of mutation for the central nucleotide (as above). Reverse complement triplets and mutations were

considered to be the same (e.g. TTT>C and AAA>G are considered to be equivalent). The locations of mutations in cancer genomes were converted from the NCBI36 genome to the GRCh37 genome for comparison with divergence data via the 'convert' tool on the UCSC genome browser. The numbers of mutations per MB were tallied and regions for which we had no human-chimpanzee divergence data were removed. Chromosome mutation rates were calculated by dividing the number of mutations found in each chromosome by the total number of valid bases (A/T/C/G) in the NCBI36 reference sequence for each chromosome.

5.3.3 Context Effects

To investigate the variance in the mutation rate associated with adjacent and adjacent-but-one nucleotides we proceeded as follows. We calculated the mutation rate for all pentamers; e.g. for TTTTT we tabulated the number of times the central nucleotide had undergone each type of mutation and divided this by the total number of TTTTTs surveyed; we also calculated the sampling variance associated with this rate as the number of TTTTTs with a mutation divided by the square of the number of TTTTTs, i.e. we assume that the number of TTTTTs with a mutation is a Poisson variate. We then calculated the variance in the mutation rate for each triplet-mutation combination. For example, if we are interested in the variance associated with the adjacent-but-one nucleotides, we calculate mean and variance in the mutation rate for XTTTY>N, where X, Y and N can be any nucleotide; for adjacent nucleotides we would consider TXTYT>N. We performed this calculation for each triplet-mutation combination. We normalized the variance associated with each triplet-mutation combination by dividing it by the square of the mean mutation rate for the triplet-mutation combination – in

essence we are dividing the mutation rate of each of the pentamers, by the mean rate across pentamers that contain the relevant triplet (e.g. all pentamers with TTT in the middle). We also calculated the average normalized sampling variance, which is subtracted from the total variance to yield an estimate of the systematic variance associated with the adjacent, or adjacent-but-one nucleotides for a particular triplet. We then averaged this estimate of the systematic variance across triplets-mutation classes to yield an estimate of the average variance associated with either adjacent or adjacent-but-one nucleotides.

5.3.4 Coincident SNPs

We calculated whether there is an excess of human and chimpanzee coincident SNPs that occur at the same position as cancer mutations as follows. We obtained the genomic locations of coincident SNPs from previous work (chapter 2) and extracted flanking sequence in each case. We then retained only those sequences where a cancer mutation was present within 500bp either side of the central coincident SNP; this left us with 603 1001bp alignments. The expected number of sites that contain both a coincident SNP and a cancer mutation was calculated by the same method as described in chapter 2, incorporating the effects of neighbouring nucleotides on the mutation rate.

5.3.5 Genomic features

To investigate what factors might influence the distribution of cancer mutations we considered whether the density per MB was correlated to a number of variables. Genomic data on the locations of telomeres and centromeres, GC content, gene density

and nucleosome association were downloaded from the UCSC website (http://genome.ucsc.edu/). Gene density was calculated as the number of nucleotides present within an exon, scaled up to the number of nucleotides per MB. We used A365 values to study the influence of nucleosome occupancy on the distribution of cancer mutation across the genome; although these values come from a skin melanoma cell line, they are highly correlated at the 1MB scale with scores from mammary cells (Dennis and MEC scores) (A365 and Dennis, $r = 0.971$, $p<0.01$; A365 and MEC, $r = -0.932$, $p<0.01$). Recombination rates per MB were obtained from Kong *et al.* (2002). As some genomic features were only available with reference to the NCBI36 genome, the divergence analysis (see above) was repeated using human NCBI36 sequences for use in the comparisons.

To compile a list of genes present in hypermutable regions, we downloaded entrez gene data from the UCSC website for each specific region. We then checked each gene against the OMIM database (http://www.ncbi.nlm.nih.gov/Omim/) for cancer association; details listed in the paper are taken directly from the OMIM website.

5.4 Results

5.4.1 The pattern of mutation

We used data of single nucleotide variants (SNVs) from each cancer genome; these are differences between the cancer genome and the reference human genome. From these we removed any variants found in the genome of the normal, non-cancerous cells from

the same individual, along with the any SNVs that matched SNPs in public databases. We removed SNPs in public databases, as other have done, to mitigate against the possibility that the individual carries an SNV that is a SNP segregating in the human population, which is not detected in the normal tissue (i.e. a false negative) but is called in the cancer tissue. We assume the remaining SNVs are new somatic mutations that have accumulated in the cancer cell lineage, though many may have accumulated before the tissue became cancerous. This left 22,910 SNVs in the SCLC genome, 33,345 SNVs in the skin cancer genome, and 31,650 and 36,641 SNVs in the AML1 and AML2 genomes respectively.

The rates of mutation for each cancer type and for the germ-line are shown in figure 5.1. In the germ-line, transitions are elevated relative to transversions as expected, with all types of transversions mutating at a similar rate. As has been mentioned before (Pleasance et al. 2009b), the pattern of mutation in the SCLC genome is dominated by G/C->T/A transversions, which are thought to be a consequence of tobacco smoke carcinogens (Pfeifer et al. 2002). As a result, the correlation between mutation rates in the SCLC genome and those observed in the germ-line is not significant ($r=0.32$, $p=0.54$); however if G/C->T/A transversions are removed the rates become comparable to those observed in the germ-line, with transitions occurring more often than transversions, and the correlation between SCLC and germ-line mutation rates becomes significant ($r=0.912$, $p=0.01$). Surprisingly, one of the AML genomes (AML1) also shows a strong excess of G/C->T/A transversions, suggesting that this leukemia might have been caused by a similar mutagen to the lung cancer. Again, the correlation between AML1 and germ-line mutation rates is non-significant ($r=0.031$, $p=0.954$); however if G/C->T/A transversions are removed the mutation rates become similar to

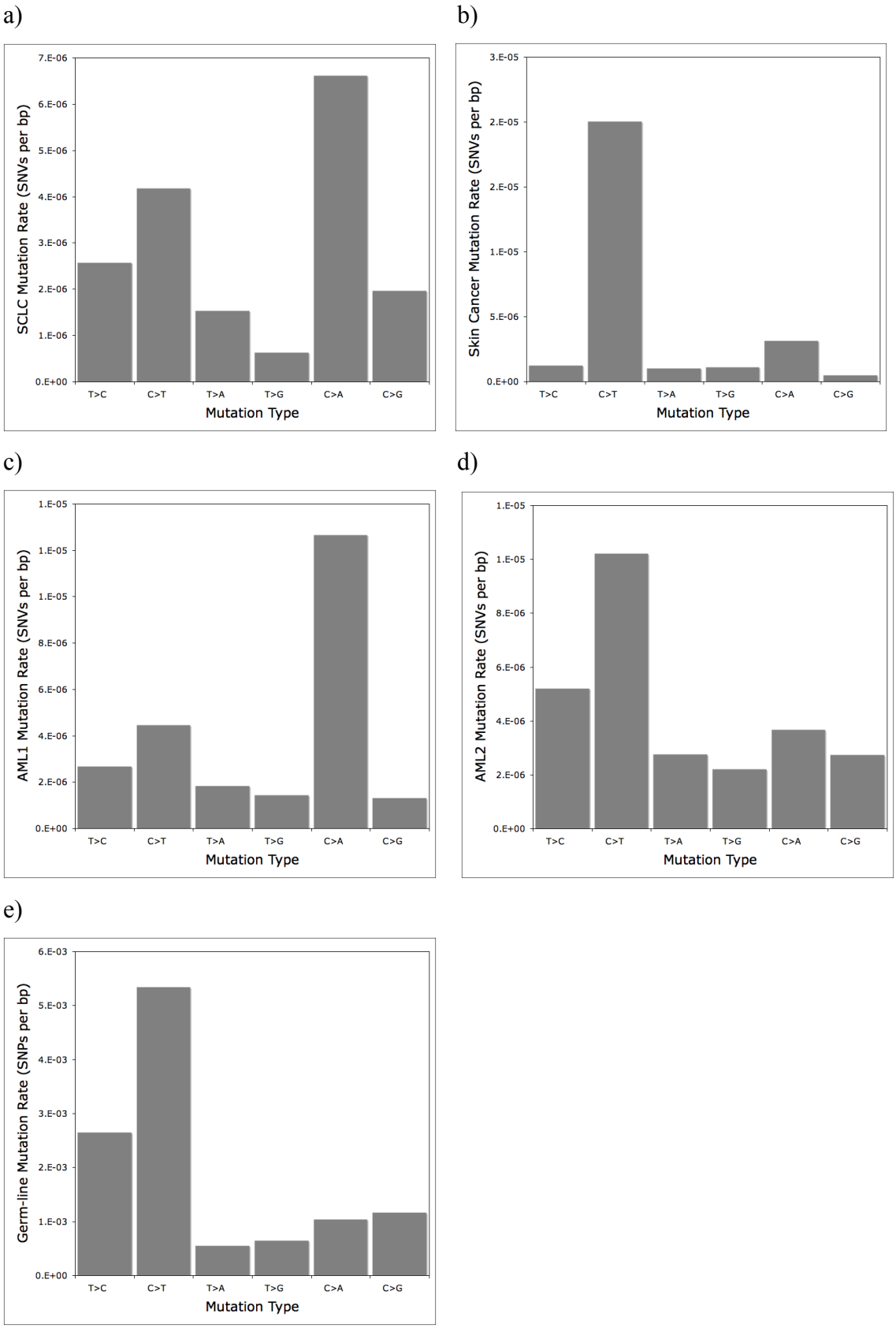


Figure 5.1. Single nucleotide mutation rates for a) SCLC, b) Skin cancer, c) AML1, d) AML2 and e) the germ-line. Transitions (C>T and T>C) are shown in the two left hand columns and transversions are shown in the four right hand columns.

those in the germ-line and the r-value of the correlation increases to 0.967, which is significant ($p<0.01$). In contrast, mutation rates in the AML2 genome strongly correlate with germ-line mutation rates ($r=0.981$, $p<0.01$). The correlation between skin cancer mutation rates and the germ-line is also significant ($r=0.905$, $p=0.01$), however this appears to be driven by an excess of G/C->A/T transitions in both genomes, which is massively over-represented in skin cancer most likely as a consequence of UV light. If G/C->A/T transitions are removed the correlation becomes non-significant ($r=-0.024$, $p=969$), and it appears that G/C->T/A transversions are over-represented in skin cancer, possibly due to the effects of reactive oxygen species at CCN and NCA triplets. T/A->C/G transitions are also under-represented in the skin cancer genome as they occur at similar rates to transversions.

The above results imply that the mutational processes occurring in AML2 cells are very similar to those in the germ-line on this scale, whereas the specific effects of UV light and reactive oxygen species have an effect on mutation rates in skin cancer.

Furthermore, the mutation patterns in SCLC and AML1 are very similar, suggesting that they may be caused by the same, or a similar, carcinogen. Indeed, partial correlations between the mutation rates of different cancers are all non-significant after controlling for germ-line rates, except for between SCLC and AML1 (table 5.1).

	SCLC	Skin cancer	AML1	AML2
SCLC		0.246	0.981*	0.617
Skin cancer			0.306	0.705
AML1				0.657
AML2				

Table 5.1. Partial correlation coefficients between different cancer genomes for single nucleotide mutation rates whilst controlling for germ-line mutation rates. * shows significance at the 5% level.

5.4.2 Local context effects of cancer mutations

The identities of the adjacent nucleotides are known to influence the mutation rate in the germ-line (Hess, Blake, and Blake 1994; Zhao and Boerwinkle 2002; Hwang and Green 2004). To investigate whether similar patterns are evident in cancer genomes we estimated the rate of each mutation as a function of the adjacent nucleotides and compared them to germ-line rates. However, since effects at single nucleotides dominate many mutations, we normalized triplet mutation rates (where the central nucleotide is the one that mutates) by dividing by the corresponding single nucleotide rate and thus focused on the residual effects associated with context. For example, the normalized mutation rate for the triplet TTT->TCT was found by dividing the triple rate by the T/A->C/G mutation rate. Furthermore, we also take the log value of each normalized mutation rate for comparisons between cancer cells and the germ-line, in order to reduce the effects of CpG transitions that mutate at a much higher rate than other triplets.

We do not find a significant correlation between the neighbouring nucleotide effects in the skin cancer mutations and those observed in the germ-line ($r=0.041$, $p=0.694$). This is not surprising since it is known that the mutagenic processes causing skin cancer have strong context effects; UV light generates pyrimidine dimers and reactive oxygen species target CCN->CAN and NCA->NAA triplets (Pfeifer, You, and Besaratinia 2005). However, when C/G->T/A transitions occurring at dipyrimidines either on the 5' or 3' side are removed, along with mutations associated with reactive oxygen species, the correlation between skin cancer and germ-line mutation rates is still non-significant ($r=-0.020$, $p=0.864$). In the previous section it was noted that both G/C->T/A

transversions and G/C->A/T transitions are over-represented in skin cancer cells, and it is only when these mutation classes are removed that the correlation between skin cancer and germ-line mutation rates becomes significant ($r=0.257$, $p=0.040$). Furthermore, T/A->C/G transitions were found to be under-represented in skin cancer cells, and if these mutations are removed, together with G/C->T/A transversions and G/C->A/T transitions, the correlation between skin cancer and germ-line mutation rates becomes much stronger ($r=0.467$, $p=0.001$). This indicates that the effects of adjacent nucleotides in skin cancer cells are much more far-reaching than just the effects associated with UV light and reactive oxygen species. In contrast, SCLC, AML1 and AML2 triplet mutation rates all correlate significantly with germ-line rates across all triplets (SCLC: $r=0.716$, $p<0.01$; AML1: $r=0.650$, $p<0.01$; AML2: $r=0.746$, $p<0.01$), however there are some obvious outliers. The rate of GCG>GGG and CTA>CAA mutations are much higher in the SCLC genome than expected from germ-line rates, whereas TCT>TTT and ACA>AAA show an increased rate of mutation in AML1 and AML2. Interestingly this effect in AML1 is not observed in SCLC suggesting that AML1 may have a similar, but not identical molecular origin. It is not known what might be driving these effects.

Although triplet mutation rates for three cancer genomes correlate significantly with germ-line mutation rates after accounting for single mutation rates, there are many similarities between cancer genomes. We performed partial correlations between each pair of cancer genomes whilst controlling for germ-line rates and found significant correlations in all but one case (SCLC and skin cancer) (table 5.2). This implies that although germ-line patterns may be repeated in somatic tissues, there are also some

	SCLC	Skin cancer	AML1	AML2
SCLC		0.177	0.380*	0.428*
Skin cancer			0.417*	0.373*
AML1				0.745*
AML2				

Table 5.2. Partial correlation coefficients between different cancer genomes for log triplet mutation rates whilst controlling for log germ-line mutation rates. * shows significance at the 5% level.

features that are unique to somatic/cancer cells, which result in different mutational signatures.

As well as neighbouring nucleotide effects, it is known that the identity of more distant nucleotides affect the mutation rate at a particular site in normal cells, even as far as 200bp away (Zhao and Boerwinkle 2002; Elango et al. 2008). To investigate whether cancer mutations show similar patterns, we estimated the variance in the mutation rate associated with the adjacent and adjacent-but-one-nucleotides, controlling for the influence of the other, for each triplet (e.g. comparing the variance in the rate associated with TNCNT with NTCTN). In all cases we found the ratio of the variance associated with the adjacent nucleotide relative to that associated with the adjacent-but-one to be significantly greater than one ($p<0.01$ for all cases). In SCLC, the average ratio of the variances is 2.78, for skin cancer it is 3.20, and for AML1 and AML2 it is 1.82 and 2.12 respectively, i.e. the adjacent nucleotides generate roughly 2-3 fold more variance in rate than the adjacent-but-one nucleotides. These values are all roughly consistent with those observed in the germ-line (Zhao and Boerwinkle 2002).

5.4.3 Coincident SNPs

It has previously been shown that there is an excess of SNPs in the human genome that also contain a SNP at the orthologous position in chimpanzee, and it has been inferred from this that some sites undergo mutation at higher rates than can be explained by the effects of adjacent nucleotides (chapter 2). To test whether this might be the cause of some mutations in cancer genomes, we compared the positions of cancer mutations from all cancer types with the locations of human and chimpanzee coincident SNPs; this

resulted in 603 alignments of 1001bp with the human-chimpanzee coincident SNP present in the central position and the cancer mutation elsewhere within the alignment. Of these alignments, one has a coincident SNP and a cancer mutation in the same position, compared to the 1.28 we would expect taking into account the effects of adjacent nucleotides on the mutation rate. As a consequence, it appears, on the basis of this very limited evidence, that the process driving the excess of human and chimpanzee coincident SNPs is not causing cancer mutations.

5.4.4 Genomic distribution of cancer mutations

To investigate the genomic distribution of cancer mutations we tallied the number of mutations per megabase (MB) for each dataset and compared the distributions to those found between human and chimpanzee genomes, which we take to be the pattern found in the germ-line, averaged over the divergence of the two species. On average there were 7.85, 11.55, 10.94 and 12.55 mutations per MB in the SCLC, skin cancer, AML1 and AML2 genomes respectively. If cancer mutations are randomly distributed across the genome we would expect the number per MB to be Poisson distributed and have a variance equal to the mean. However, in all cases the variance is significantly greater than the mean and so the number of mutations per MB is significantly over-dispersed (the observed variances are as follows: SCLC = 37.09, $p<0.01$; skin cancer = 53.26, $p<0.01$; AML1 = 76.40, $p<0.01$; AML2 = 1273.81, $p<0.01$). The density of mutations in each of the cancer genomes is weakly but significantly correlated to that found in the germ-line (SCLC: $r=0.129$, $p<0.01$; skin cancer: $r=0.195$, $p<0.01$; AML1: $r=0.195$, $p<0.01$; AML2: $r=0.335$, $p<0.01$); the correlation is of very similar strength for all cancers except AML2, which shows a stronger correlation. As expected, given the

rather weak correlations we observe, there is significant variation in the density of mutations in cancer genomes which cannot be attributed to the germ-line pattern; if we randomly distribute the same number of mutations observed in each cancer across the genome, weighted by the density of mutations per MB observed between human and chimps, we find the observed variance to be significantly greater than in the simulated data ($p<0.01$ for all cancers).

There are also significant partial correlations between the densities of mutations in the different cancer genomes after controlling for human and chimp divergence rates (table 5.3). For most comparisons these correlations are weak, except between the two AML genomes and between SCLC and skin cancer. Most of the correlation between the two AML genomes can be explained by high rates of mutation around centromeres, some telomeres and five other 1MB regions (discussed later). If these MBs are removed the r-value is reduced to a modest 0.101, which is still significant ($p<0.001$), but explains little of the variance in either dataset. In the case of SCLC and the skin cancer, there are no obvious outliers. This implies that many patterns are consistent across the two different cancers on this scale, even though the effects of specific carcinogens vary.

To investigate the distribution of cancer mutations in more detail we compared the density per MB with a number of key genomic features using step-wise multiple regression, retaining only those features that contributed significantly to the regression model. The features investigated were the distance to the telomere, the distance to the centromere, the GC content, the human and chimpanzee divergence, the gene density, the recombination rate and the nucleosome association rate (table 5.4). The multiple regression model for SCLC explains approximately 24% of the total variance in the

	SCLC	Skin cancer	AML1	AML2
SCLC		0.399*	0.024	-0.101*
Skin cancer			0.105*	0.001
AML1				0.674*
AML2				

Table 5.3. Partial correlation coefficients between different cancer genomes for the frequency of mutations per MB whilst controlling for human and chimp divergence. * shows significance at the 5% level.

Feature	**SCLC**	**Skin Cancer**	**AML1**	**AML2**
Telomere	-0.092	NS	-0.039	NS
Centromere	NS	NS	-0.052	-0.065
GC content	-0.307	0.050	-0.059	NS
Divergence	0.170	0.279	0.193	0.262
Gene Density	-0.048	-0.071	NS	-0.039
Recombination	NS	-0.059	-0.048	-0.139
Nucleosome Association	NS	-0.101	NS	0.124
r	0.490	0.355	0.216	0.311
p	<0.01	<0.01	<0.01	<0.01

Table 5.4. Multiple regression analysis of key genomic feature for each cancer type. Features used are distance to the telomere (telomere), distance to the centromere (centromere), GC content, Human and Chimp divergence rates (Divergence), gene density, recombination rate (recombination) and nucleosome association rate (nucleosome association). Partial correlations whilst controlling for all other significant features are shown for each genomic feature.

distribution of SCLC mutations, whereas this figure is 21.7%, 4.7% and 11% for skin cancer, AML1 and AML2 respectively. Taking the data together, it appears there are many features that may explain at least part of the distribution of cancer mutations, however in many cases the low partial correlation scores suggest that some features have limited explanatory power. First, as expected given the correlations above, human and chimp divergence is included within the regression model for all cancer types, and partial correlations suggest that it has a significant impact on the distribution of cancer mutations. This implies that some regions in the human genome are more mutable than others, and that this remains true in cancer genomes regardless of other mutational pressures. Second, gene density is included in the model for all cancers except AML1; however, although the effect is significant in all the other cancers it has little predictive power by itself. The partial correlations are negative in all cases, showing that there are less cancer mutations in genic regions. This is probably due to a number of factors – the higher rate of transcription-linked repair processes, as highlighted in some of the original cancer genome papers (Pleasance et al. 2009a; Pleasance et al. 2009b), and negative selection. Third, cancer mutations are negatively correlated with recombination rates in the SCLC, AML1 and AML2 genomes, making a significant contribution to the predictive power of the regression model, particularly in the case of AML2. As cancer mutations occur in somatic tissues it is highly unlikely that the recombination process itself has a direct impact on mutagenesis since somatic recombination is thought to be very rare (Paques and Haber 1999). However, the correlation may be driven by other underlying factors that have not been considered here. All other features have either a non-significant, limited or contradictory (i.e. they differ between cancers) contribution to the regression model in each case and across all cancers. For example, there is a significant negative partial correlation between SCLC

and GC content when controlling for all other features in the final regression model, but this is only mirrored in part in AML1, not at all in AML2 and is positively correlated in skin cancer. This result may reflect the differing causes of each type of cancer. Furthermore, nucleosome association rate is strongly negatively correlated to skin cancer mutation density and strongly positively correlated to AML2 mutation density. This might not seem surprising, since the nucleosome binding scores come from a skin melanoma cell line; however, nucleosome binding scores are highly correlated between the melanoma cell line and a mammary gland cell line (see methods). Finally, a prominent feature not picked up by the regression analysis is the propensity of AML mutations to be located immediately adjacent to centromeres. Although distance to the centromere is included in the regression analysis, it is not a significant indicator as the mutation density does not decline with distance from the centromere outside of the first few MBs. However, the effect is very strong with 11.8% and 27.8% of mutations occurring within the 2.58% and 2.37% of sequence that immediately flanks the centromeres in AML1 and AML2 respectively.

5.4.5 Outlier regions and implications for cancer

There are some megabase regions in cancer genomes that contain a very high density of mutations. Appendix 5.1 contains a list of regions that fall outside of four standard deviations from the mean (on the order of 0.01% of the data assuming a normal distribution), together with any genes located in those regions for each cancer. As the AML genomes are dominated by mutations in centromeric and telomeric regions, these have been removed before outliers are identified, and all functional gene information in the following section was taken from the OMIM website

(http://www.ncbi.nlm.nih.gov/omim). In the SCLC genome none of the 1MB regions with high numbers of mutations overlap with the other three datasets outside of the centromeric regions. Within the SCLC outliers there are two regions that stand out: chromosome X: 61MB-62MB and chromosome 6: 57MB-58MB, both of which are located close to the centromere. The region on the X contains no genes whereas the region on 6 contains four genes. Furthermore, 42 of the mutations from chromosome 6: 57MB-58MB fall within a 100KB region just downstream of PRIM2, which is involved in DNA replication. This is a region we have previously shown to contain a high density of sites at which both humans and chimpanzees have a SNP (chapter 3). In the skin cancer genome there are again no highly mutated regions that overlap with other cancer genomes outside of those close to centromeres. There are, however, two regions that contain genes that are associated with cancer; genomic alterations in GPC5 have been implicated in lymphomas, lung cancers and squamous cell carcinoma, whereas loss-of-function in TLR4 has been linked with more rapid relapse after treatment for breast cancer. In the two AML genomes there are five 1MB regions that have high rates of mutation in both AML1 and AML2. Furthermore, the high numbers of mutations within these regions actually fall within smaller areas of 100-200KB, which are consistent across both genomes. Within these five regions there are many genes linked to cancer. In one MB region in the MHC locus there is AGER, which codes for a receptor that has been shown to decrease growth and metastases of cancer cells when blocked in mice, HLA-DRA, which is linked with thromboembolic complications in cancer patients, and TAP1, which was found to contain a mutation in SCLC cell lines. In other highly mutable regions of the AML genomes, PRR4 has been linked to breast cancer, ETV6 is a known tumour suppressor gene and LYPD1 is a tumour suppressor gene that has a role in triggering apoptosis. Furthermore, highly mutable regions

unique to either AML genome also contain cancer genes. In AML1, MASL1 is related to the oncogene MAS1, and HLA-G has been shown to have a high level of transcription in malignant melanoma cells. In AML2, MMP21 was detected at unusual levels in cancer cells, BCCIP is known to interact with the breast cancer gene BRCA1 and has been shown to inhibit growth of some breast and brain tumour cells, whereas HIST2H4A, which codes for histone H4, was found to be lost in monoacetylated and trimethylated forms in cancer cells.

The occurrence of many cancer genes in regions that are undergoing high rates of mutation in cancer genomes may suggest that some genes are involved in disease phenotypes simply because they mutate at higher rates. To test this we downloaded a census of human cancer genes from the Cancer Genome Project (http://www.sanger.ac.uk/genetics/CGP) that was originally compiled by Futreal *et al.* (2004). Genes were only included in this list if they have been causally implicated in oncogenesis, have mutations that have been presented in at least two independent reports from primary patient material and are mutated in areas outside of methylated promoter regions. Changes in expression level alone are not sufficient for genes to be included in this dataset. For each cancer genome we then tested for a correlation between the number of cancer mutations per MB and whether there was a gene implicated in that specific cancer present in the MB using a logistic regression. In all four cases the result was non-significant (SCLC: $p=0.31$, skin cancer: $p=0.46$, AML1: $p=0.31$, AML2: $p=0.414$), and results are similar using t-tests. However, it may be the case that many regions in the genome cannot be causally implicated in cancer as they contain either no genes, or genes that are not part of pathways that can lead to cancer progression. As a result, we repeated the analysis as before, considering the number of

mutations and whether there were genes specific to each cancer type in each region, however in this case we only included those regions that contain genes associated with at least one type of cancer; we again found the result to be non-significant in all cases (SCLC: p=0.90, skin cancer: p=0.25, AML1: p=0.94, AML2: p=0.50), with similar results obtained from t-tests. This suggests that particular genes are not repeatedly associated with cancer simply because they have a high mutation rate. However, since the genetic complexity and full array of causative cancer genes are not yet understood, the results cannot be assumed to be definitive.

5.4.6 Chromosomal mutation rates

There is significant variation in chromosomal mutation rates across all four datasets, which can be seen in figures 5.2a-d as not all 95% confidence intervals overlap for each chromosome. This is consistent with the pattern of germ-line mutation inferred from the divergence between human and chimpanzee (Mikkelsen et al. 2005). For SCLC, skin cancer and AML1, the mutation rates of the sex chromosomes fall within the confidence intervals for the mutation rate of at least one autosome. This is markedly different from human and chimpanzee divergence, where the Y chromosome mutates at a higher rate than the autosomes, and the X chromosome at a lower rate (Mikkelsen et al. 2005). However, this is not unexpected as cancer mutations in these datasets are occurring in somatic cells, so all chromosomes undergo the same number of replications. This is in contrast to events in the germ-line where the Y chromosome undergoes more mutations than the autosomes, which undergo more mutations than the X chromosome due to relative time spent in males that undergo more cell divisions (Haldane 1947). Conversely, for AML2 the mutation rate on the Y chromosome is

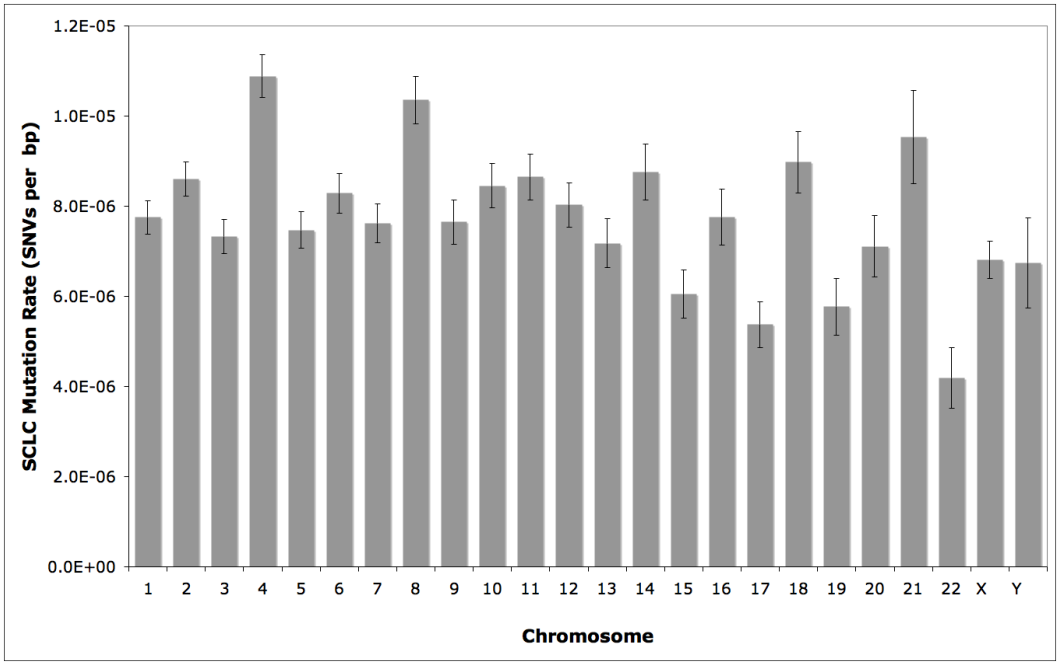


Figure 5.2a. Chromosome mutation rates for the SCLC genome. Error bars represent 95% confidence intervals.

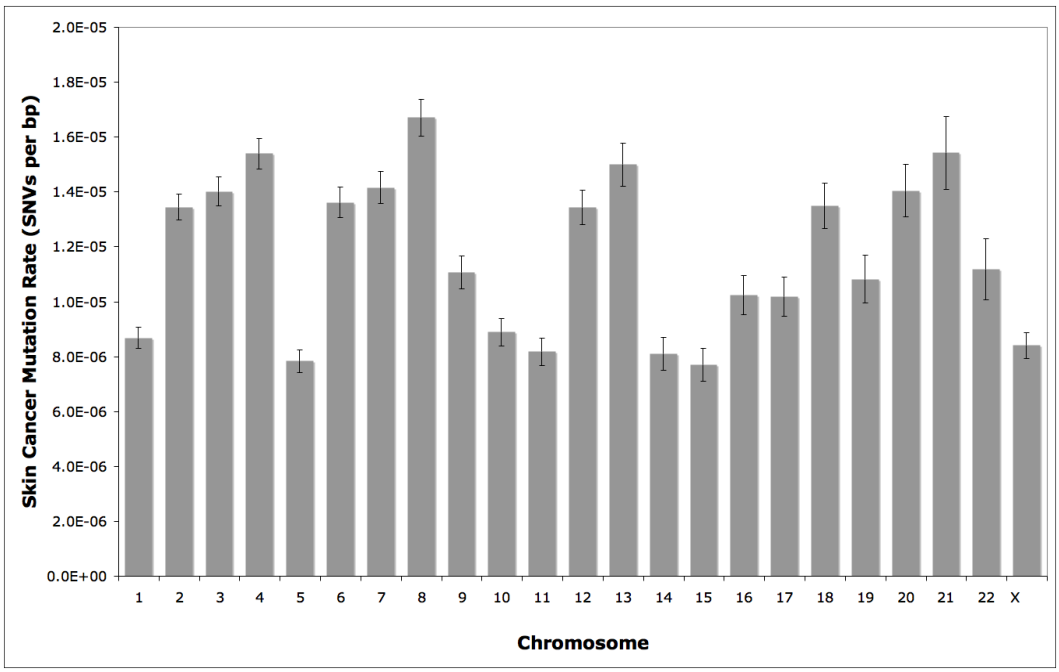


Figure 5.2b. Chromosome mutation rates for the skin cancer genome. Error bars represent 95% confidence intervals.

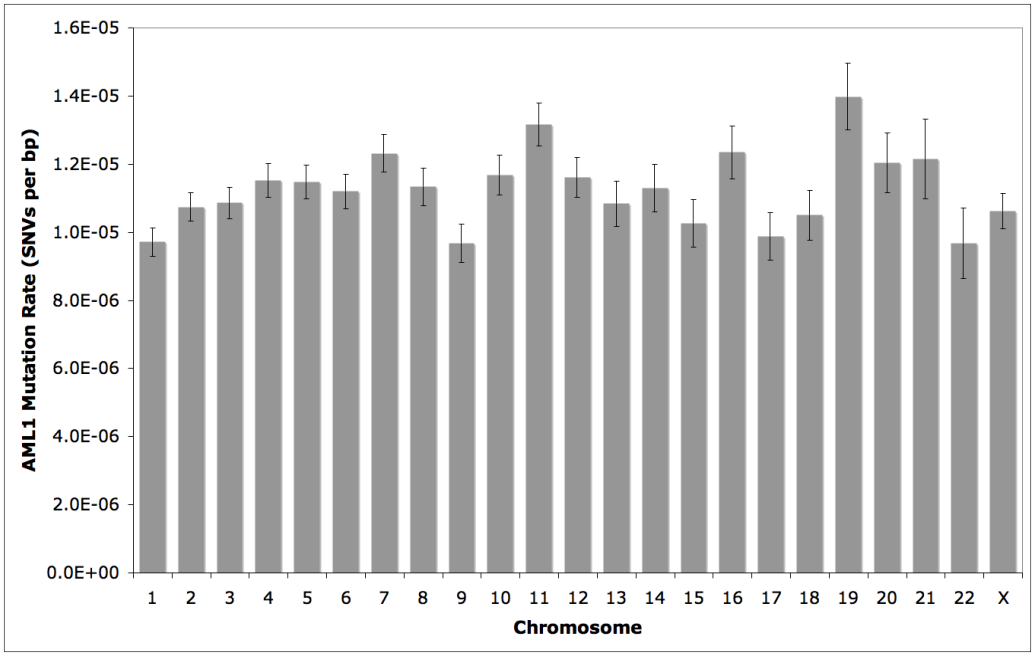


Figure 5.2c. Chromosome mutation rates for the AML1 genome. Error bars represent 95% confidence intervals.

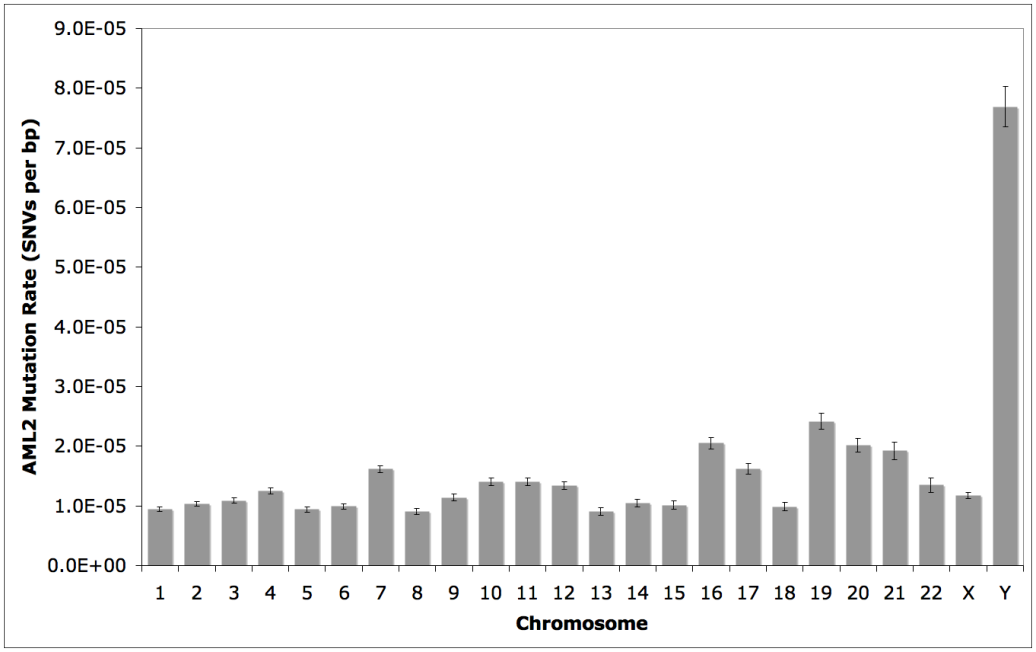


Figure 5.2d. Chromosome mutation rates for the AML2 genome. Error bars represent 95% confidence intervals.

	SCLC	Skin cancer	AML1	AML2	Divergence
SCLC		0.406	0.230	-0.261	-0.092
Skin cancer			0.068	0.017	0.115
AML1				0.644*	0.285
AML2					0.520*
Divergence					

Table 5.5. Correlation coefficients between different cancer genomes for the frequency of mutations per chromosome. * shows significance at the 5% level.

significantly higher than the autosomes and X chromosome, which is not easily explained given the somatic nature of the mutations. Furthermore, there are only two significant correlations when we compare the autosomal mutation rates of the four cancer genomes, and the germ-line rates (table 5.5). Firstly, AML2 correlates significantly with the germ-line rates, which is perhaps not surprising given that they also correlation significantly on the MB scale, and for the most part, show similarities in mutation patterns when considering neighbouring nucleotides and single nucleotides. Secondly, AML1 chromosome mutation rates correlate significantly with AML2 rates, which remains significant after controlling for human and chimp divergence rates ($r=0.605$, $p=0.004$). Again, this is not surprising since there is a strong correlation on the MB scale and for neighbouring nucleotide rates.

5.5 Discussion

The recent sequencing of four cancer genomes provides a unique opportunity to study the patterns of mutation on various scales occurring in different cancers (SCLC, skin cancer and AML in this case), but also within a type of cancer (AML). We observe some striking similarities and differences between the cancer genomes, and between the cancer genomes and the germ-line at the three scales we have considered. It is important to appreciate that given just four cancer genomes it is currently not possible to determine whether the differences we observe represent systematic differences between cancers and the somatic tissue they developed from, or random differences between individuals. However, the analysis of the AML genomes suggests that the differences between the cancer genomes are likely to be due systematic differences

rather than random differences between individuals, since these genomes show strong correlations in their patterns of mutation at all scales, differing only in the high rate of C>A mutations in AML1.

We assume, as others have done, that most mutations are 'passengers' effectively hitchhiking along with a small handful of 'driver' mutations that have a causal relationship with cancer (Greenman et al. 2007). In fact, although it has been suggested that one of the steps in the development of cancer might be the loss of some DNA repair enzymes and hence the increase in the mutation rate (Loeb, Bielas, and Beckman 2008), it seems likely that the vast majority of mutations that are observed in the cancers tissue accumulated prior to the development of cancer. The rate of somatic mutation has been estimated to be 0.77×10^{-9} per site per cell division, and this is expected to generate of the order of 10,000s of mutations in dividing cells in a mature adult (Lynch 2010).

At the single and neighbouring nucleotide scale, mutations are often a consequence of specific carcinogens associated with different cancer types. Mutations in SCLC bear the hallmarks of the mutagens present in tobacco smoke, with an over-representation of C>A mutations, whilst mutations in the skin cancer genome are often at dipyrimidines, as one would expect from UV radiation, or show the patterns associated with reactive oxygen species. However, there also appears to be more variation in the rate of different types of mutation in the skin cancer cell above that associated with specific carcinogens. Mutations in AML are different; in AML1 there is an over-representation of C>A mutations, whereas in AML2, most mutation rates are similar to those found in the germline, implying that mutation patterns in the two different AML genomes – both of subtype M1 – may have different causes. Intriguingly, the patterns of mutation in the

AML1 genome are very similar to those observed in the SCLC genome. The most obvious interpretation of this is that the mutation patterns in the AML1 and SCLC genomes are being caused by the same carcinogens, namely those found present in tobacco smoke. Tobacco smoke has been linked to many types of cancer (Doll 1996; Kuper, Boffetta, and Adami 2002), and some studies have shown a significant, albeit weak, association between cigarette smoking and myeloid leukemias (Kinlen and Rogot 1988; Wakabayashi et al. 1994). However, there are also studies showing no significant link (Adami et al. 1998; Stagnaro et al. 2001), making the association not entirely clear. Alternatively, it could be that very similar carcinogens are involved in both cases, rather than the effects being specific to tobacco smoke. The comparable mutation patterns could also be enhanced as a consequence of similar deficiencies in replication and repair enzymes in the two cancer cells; in this case this could involve Y family polymerases that are implemented in cases where bulky adducts attach to DNA nucleotides (Loeb and Monnat 2008). Unfortunately the smoking history of the individual from which AML1 was taken is unknown, so the cause of cancer in this case remains unidentified.

The influence of neighbouring nucleotides appears to be fairly consistent across all cancer genomes, except for the skin cancer genome, which shows no similarity to the germ-line. It seems likely that neighbouring nucleotide effects are a consequence of biases in the replication or repair machinery, so the consistency across tissues is possibly not surprising.

On a larger scale there is significant variation in the numbers of mutations per megabase above that expected given human and chimp divergence rates, and some regions appear to be hyper-mutable. The specific effects of carcinogens do not appear to explain this

pattern, since the rate of mutation at the 1MB scale is only weakly correlated to GC content in two of the cancer genomes, and in the two cancers in which the correlation is at all strong (SCLC and skin cancer) one might expect the correlation to be in the opposite direction (SCLC: $r=-0.350$, $p<0.01$, Skin cancer: $r=-0.352$, $p<0.01$; AML1: $r=-0.069$, $p<0.01$; AML2: $r=0.003$, $p=0.861$). Furthermore, across the three different types of cancer, 1MB regions with very high mutation rates appear to occur in different regions of the genome. However, the two AML genomes share many mutation hotspots, mostly at centromeres and telomeres, but also at five other 1MB locations in the genome. In addition, within the five regions that do not fall near to the centromere or telomere, there are even smaller 100-200kb regions that contain the majority of the mutations that are the same in both AML1 and AML2. It seems unlikely that these have arisen by chance, thus implying a causative process at this scale. Note that we do not have enough mutations per MB to explore the density in mutations at a finer scale than 1MB, however further cancer genomes will allow us to do this. Several of the regions with the highest mutation rates contain a number of genes that have been linked to some form of cancer in the literature, suggesting that some genes may be associated with cancer simply because they are in mutation hotspots and are therefore more likely to mutate. However, we find no evidence that regions containing genes specific to each type of cancer have higher mutation rates than other regions in the genome, even amongst regions containing genes associated with other cancers.

When considering all 1MB regions we see at least marginally significant correlations between most cancers, possibly driven by background mutations, and indeed all four cancer genomes correlate significantly with human and chimp divergence on the same scale. This points to the idea that impaired replication and repair mechanisms may play

a part in all cancers as mutation patterns at the 1MB scale are similar to those expected in the germ line. Two of the correlations between cancer genomes are particularly strong; there is a strong correlation between AML1 and AML2, however when outlier regions are removed the correlation is drastically reduced, and between SCLC and skin cancer, which occurs in spite of the lack of correlation among outlier regions. It is possible that the variation in the mutation rate at the 1MB scale is a consequence of processes such as chromatin openness or nucleosome binding, and this would readily explain why the cancer types do show differences with each other and with the germ-line. However, it is not clear that chromatin state should vary wildly between different tissues (Ozsolak et al. 2007), and nucleosome occupancy scores from skin melanoma cells (A375), mammary epithelial cells (MEC) and mammary gland cells (Dennis) correlate strongly (A375 vs MEC: $r=-0.932$, $p<0.01$; A375 vs Dennis: $r=0.971$, $p<0.01$; Dennis vs MEC: $r=-0.982$, $p<0.01$).

Finally, on the chromosome level, and specifically for the autosomes, there are very little similarities between cancer types and all correlations are non-significant. Furthermore, correlations between three of the cancer genomes (SCLC, skin cancer and AML1) and human and chimp divergence are non-significant. Conversely, the chromosome mutation rates of the two AML genomes correlate significantly, and AML2 correlates significantly with human and chimp divergence, perhaps reinforcing the notion that most mutations in the AML2 genome are occurring in a similar way to that expected in normal cells.

The genetic basis of cancer appears to be very complex, and there is still much that is unknown. Here, we show that although there is significant variation between rates of

mutation on the local, 1MB and chromosomal scale within each genome, there are also similarities between different cancer types. However, this picture may be muddied further if there is also significant heterogeneity between cancer cells in a single individual (Fox, Salk, and Loeb 2009).

Acknowledgements. AH and AEW were funded by the BBSRC. Funding support for the AML Sequencing Project was provided by a gift from Alvin J. Siteman, and from grants from the NCI (CA101937), NHGRI (U54HG003079), and the Barnes-Jewish Foundation.

6. Discussion and Conclusions

6.1 The impact of mutations

The rate of mutation in the human genome is estimated to be ~2 x 10^{-8} per site per generation (Nachman and Crowell 2000; Kondrashov 2003; Xue et al. 2009), although it is clear that this rate varies considerably over a number of different scales and contexts (Ellegren, Smith, and Webster 2003). The importance of understanding this variation is far reaching; mutations can result in deleterious effects at certain sites and in certain regions of the genome, possibly causing genetic diseases or cancer. Appreciating which particular contexts and regions are likely to undergo more mutation events may lead to a greater diagnostic ability and an understanding of why certain genes and mutations are involved in disease. Furthermore, an understanding of variation in the mutation rate can aid in the description of how and why species evolve; the use of mutation rates as molecular clocks and the subsequent analysis of phylogenetic trees through comparative sequence analysis, together with an understanding of diversity levels within populations, shed much light on our history. To correctly infer phylogeny and particularly the timescale over which evolution occurs, it is necessary to understand variation in the mutation rate. Considering variation in the mutation rate is also important when unlocking the secrets of the genome; mutation patterns are key to identifying functional regions, understanding why other regions are conserved and quantifying processes such as recombination and repair. In this thesis we considered further patterns of sequence variation that cannot be explained with current knowledge of the processes that drive differing mutation rates. This has led us to a general conclusion that there is much more variation in the mutation rate in the human

genome, both in the germ-line and the somatic tissues, than was thought. Here, I will summarise the results obtained in this thesis, highlighting the broader issues that emerge from our analyses.

6.2 Cryptic variation in the mutation rate

In chapter 2 we showed that there is more variation in the mutation rate of single sites than is currently expected, given what is known about the effects of neighbouring nucleotides on the rate of mutation; we refer to this as 'cryptic variation'. This result was inferred from the excess of sites in the human genome that also have a SNP in the orthologous position in chimpanzee above that expected at random. As we controlled for local context as well as other known mutational patterns and natural selection, it seems that the causes of this variation in the mutation rate could not be explained by previous knowledge. However, we must be cautious in quantifying the impact of cryptic variation in the mutation rate since we were unable to rule out a contribution to the excess of coincident SNPs from ancestral polymorphisms and substitutions in paralogous sequences that occurred before the split of human and chimp, although it seems unlikely that these features contribute substantially to the excess of coincident SNPs. One interesting feature of this variation is that due to the nature of its detection, it only impacts single sites in isolation, rather than particular regions of the genome, ruling out any small-scale regional effects as a cause. This is further supported by results in chapter 3 where we fail to find a strong correlation between coincident SNP densities and any other feature on the genomic scale. Outside of the specific actions of replication and repair activity, it is difficult to imagine a process that could operate in such a fashion, although DNA topology and packaging could certainly be a candidate if

specific sites are more important than others in controlling the structure of DNA. Of course, some coincident SNPs will be a consequence of known factors, and indeed there is some evidence that balancing selection may maintain a small subset of polymorphisms at the same position in both humans and chimp. However, the general pattern of excess does not resolve with the introduction of a single causative mechanism. It may well be interesting to consider whether there is an excess of coincident SNPs between different pairs of species, or indeed between more distantly related species, as this may give an indication of the evolutionary distance at which the causative process is preserved and possibly lead to greater insights into cryptic variation in the mutation rate.

The strength of cryptic variation in the mutation rate could potentially be as large as that associated with the CpG effect. It has been estimated that roughly 1/3 of all disease causing germ-line mutations are associated with increased rates of mutation at CpG sites (Cooper and Youssoufian 1988; Cooper and Krawczak 1990). It therefore may seem plausible that hypermutable sites that are associated with cryptic variation may also contribute a similar amount to genetic disease mutations. Of course, this is merely speculation as the processes that cause cryptic variation in the mutation rate may well be very different to those associated with CpG hypermutation, particularly as the latter seems to have nothing to do with replication and is instead a consequence of deamination at methylated cytosines (Coulondre et al. 1978; Bird 1980). However, since we have shown that the excess of coincident SNPs exists for both CpG and non-CpG mutations, it is difficult to rule this out. Unfortunately, coincident SNPs are not associated with any specific context, either on a local of genomic scale, and so unlike mutations occurring as a result of adjacent nucleotides, it may be very difficult to

identify particular sites that are more or less likely to undergo mutation and thus aid in clinical detection. Indeed, our only attempt thus far to link diseases with cryptic variation in the mutation rate was performed by comparing the locations of cancer mutations to human and chimpanzee coincident SNP sites, and in this case we found no excess of such events above that expected by chance. However, this does not preclude the effects of cryptic variation in the mutation rate in causing genetic disease, as the sample sizes we used were very small and do not allow a definitive conclusion. Furthermore, the process may also vary between germ-line and somatic tissues.

Cryptic variation in the mutation rate is potentially a problem in evolutionary genetics since it can affect the ability to estimate evolutionary distances and in particular to accurately infer the ancestral state; this has been found to be a problem when classifying CpG and non-CpG mutations (Gaffney and Keightley 2008).

6.3 Simultaneous mutation

In chapter 4 we showed that approximately 3% of all distinct SNPs arise as a consequence of simultaneous mutation across the two strands of a DNA helix, either in a single process, or by two tightly linked events in which the first mutation causes the second on the opposite strand. It is generally assumed that a single mutation event generates a single new allele in the population. However, we showed that there are an excess of sites in the human genome that contain three alleles, and that the minor alleles at these sites tend to cluster on phylogenetic trees of individuals in the population. It is possible that the excess of tri-allelic sites could be generated by mismatches occurring in heteroduplex DNA during recombination, however we showed that there is no

association between recombination rate and sites that have three alleles. A similar process is thought to occur between adjacent nucleotides on the same strand (Averof et al. 2000), however we estimate that this process has a smaller impact on diversity in the human population, generating slightly less that 1% of all distinct SNPs.

The importance of simultaneous mutation in the process and analysis of evolution may be far reaching. Failing to account for the generation of two new alleles in a single event may lead to mis-inferences in phylogenetic analysis of populations; minor alleles will not necessarily be grouped together using standard approaches of phylogenetics. Although this is only likely to become a problem in small SNP datasets, it is not known whether simultaneous mutation events tend to cluster in certain parts of the genome where this issue could become more important. Furthermore, it is possible that simultaneous mutation events in coding regions may allow for a more rapid exploration of evolutionary pathways than single mutations occurring sequentially. Two new alleles generated at a non-synonymous site could result in two different amino acids, the effects of which may be dramatically different.

6.4 Cancer Mutations

In chapter 5 we showed that the patterns of mutations in different types of cancer cells are complex. In general, cancer mutations show the strongest similarity to germ-line mutations when considering the effects of neighbouring nucleotides and at the 1MB scale. However, at the single nucleotide level, most of the cancers considered showed patterns of mutation associated with specific carcinogens, for example, many skin cancer mutations occurred at dipyrimidines as a consequence of UV light (Pfeifer, You,

and Besaratinia 2005). Most interestingly, one of the acute myeloid leukemia (AML) genomes showed very similar patterns of mutation to those observed in the small cell lung cancer genome. The most obvious explanation for this is that AML in this case may have been a consequence of tobacco smoke. However, as the smoking history of the patient is unknown, we cannot rule out other causes. At larger scales (1MB and whole chromosomes), there are substantial differences between the distributions of mutations in different types of cancer for the most part, although some cancers also show similarities. Perhaps the most insightful aspect of the analysis at this level is the comparison of two individuals with the same type of cancer, AML; mutations in these two genomes show the strongest correlation, perhaps implying that the differences in the patterns of mutation between cancers are likely to be systematic rather than stochastic, although we must be cautious in making firm conclusions with a limited number of datasets.

Most mutations in cancer are thought to be passengers that have no causal affect on the disease. However, it is possible that some genes are involved in cancer simply because they are in regions of the genome that mutate at higher levels than others. We performed an analysis to test this theory and found little support, however since the full array of genes associated with each type of cancer is unknown, we cannot come to any firm conclusions. It is clear that cancer in a very complex disease, with varying patterns of mutation across different cancers, but also many similarities.

6.5 Variation in the mutation rate

There have been many studies over the past 60 years that have analyzed patterns of mutation in the human genome and subsequently detailed the various ways in which a site can vary in its rate of mutation. Variation has been observed on numerous levels: at single sites, as a consequence of local context, on a regional scale and between different chromosomes. Perhaps the largest variation in the rate of mutation occurs at the smallest scale, the single nucleotide, however it is unknown the extent to which this variation is explained by processes associated with primary sequence context. Here, we have shown that there is far more variation at this scale than was previously thought. This has far reaching implications in the study of evolution and disease, since it may change the interpretation of the levels of conservation we might expect amongst homologous sequences, alter the way we look for selection, or point to hypermutable sites that are implicated in disease. Although we do not currently understand the underlying mechanisms that generate cryptic variation in the mutation rate, more data and analysis may lead to answers in the future. It would be interesting to compare coincident SNP distributions to other features of DNA structure, both within primates and across a wider range of species. Furthermore, we have shown that nucleotide changes may arise at this scale through a novel form of mutation where two new alleles are generated in a single, or tightly linked series of events. Although potential underlying mutational mechanisms that could lead to this result are well established, it has not previously been shown that they could generate variation in this way and at this frequency. The process of simultaneous mutation could be further understood with the use of Y chromosome sequence data that is obtained in an unbiased fashion; this would allow a greater understanding of the processes that cause tri-allelic SNPs. Similarly, it

may be interesting to consider whether simultaneous mutation events occur non-randomly across the genome, and what impact this might have on evolution.

We have also shown that the patterns of mutation can vary in cancer genomes beyond that observed in the germ line. This brings up an interesting question as to whether there is more variation in the mutation rate between cells from different tissues, in different individuals and under differing conditions. Although cancer genomes are likely to represent an extreme, due to the likely increased level of mutation, the data does serve to highlight that there is much more we don't know about the scale of variation in the mutation rate. The sequencing of more cancer genomes will undoubtedly lead to a greater understanding of the disease and patterns of mutation. A comparison of more individuals with the same type of cancer may allow us to assess the cause of variation, and subsequent comparison to other forms of cancer may identify important differences.

More recent advances in sequencing technology have made large sequences projects more feasible, indeed projects like the environmental genome project (www.egp.gs.washington.edu/) and the 1000 genomes project (www.1000genomes.org) will generate masses of data for further analysis. This, together with practical laboratory analysis, may finally allow the scientific community to understand the full nature of mutation and its underlying processes.

References

Adami, J. et al. 1998. Smoking and the risk of leukemia, lymphoma, and multiple myeloma (Sweden). Cancer Causes Control **9**:49-56.

Adams, A. M., and R. R. Hudson. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics **168**:1699-1712.

Antonarakis, S. E., M. Krawczak, and D. N. Cooper. 2000. Disease-causing mutations in the human genome. European Journal of Pediatrics **159**:S173-S178.

Asthana, S. et al. 2007. Widely distributed noncoding purifying selection in the human genome. Proc Natl Acad Sci U S A **104**:12410-12415.

Atamna, H., I. Cheung, and B. N. Ames. 2000. A method for detecting abasic sites in living cells: age-dependent changes in base excision repair. Proc Natl Acad Sci U S A **97**:686-691.

Averof, M. et al. 2000. Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science **287**:1283-1286.

Baer, C. F., M. M. Miyamoto, and D. R. Denver. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nature Reviews Genetics **8**:619-631.

Bailey, J. A. et al. 2002. Recent segmental duplications in the human genome. Science **297**:1003-1007.

Bailey, J. A. et al. 2001. Segmental duplications: Organization and impact within the current Human Genome Project assembly. Genome Research **11**:1005-1017.

Belle, E. M. et al. 2005. An investigation of the variation in the transition bias among various animal mitochondrial DNA. Gene **355**:58-66.

Benton, M. J., and P. C. Donoghue. 2007. Paleontological evidence to date the tree of life. Mol Biol Evol **24**:26-53.

Bird, A. P. 1980. DNA Methylation and the Frequency of Cpg in Animal DNA. Nucleic Acids Research **8**:1499-1504.

Blake, R. D., S. T. Hess, and J. Nicholsontuell. 1992. The Influence of Nearest Neighbors on the Rate and Pattern of Spontaneous Point Mutations. Journal of Molecular Evolution **34**:189-200.

Bohossian, H. B., H. Skaletsky, and D. C. Page. 2000. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. Nature **406**:622-625.

Boulikas, T. 1992. Evolutionary Consequences of Nonrandom Damage and Repair of Chromatin Domains. Journal of Molecular Evolution **35**:156-180.

Brown, T. C., and J. Jiricny. 1988. Different Base Base Mispairs Are Corrected with Different Efficiencies and Specificities in Monkey Kidney-Cells. Cell **54**:705-711.

Bubb, K. L. et al. 2006. Scan of human genome reveals no new Loci under ancient balancing selection. Genetics **173**:2165-2177.

Burgess, R., and Z. Yang. 2008. Estimation of hominoid ancestral population sizes under bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol Biol Evol **25**:1979-1994.

Casane, D. et al. 1997. Mutation pattern variation among regions of the primate genome. Journal of Molecular Evolution **45**:216-226.

Chamary, J. V., J. L. Parmley, and L. D. Hurst. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet **7**:98-108.

Chang, B. H. et al. 1994. Weak male-driven molecular evolution in rodents. Proc Natl Acad Sci U S A **91**:827-831.

Chen, C. L. et al. 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Research **20**:447-457.

Cooper, D. N., and M. Krawczak. 1990. The Mutational Spectrum of Single Base-Pair Substitutions Causing Human Genetic-Disease - Patterns and Predictions. Human Genetics **85**:55-74.

Cooper, D. N., and H. Youssoufian. 1988. The Cpg Dinucleotide and Human Genetic-Disease. Human Genetics **78**:151-155.

Coulondre, C. et al. 1978. Molecular-Basis of Base Substitution Hotspots in Escherichia-Coli. Nature **274**:775-780.

Crow, J. F. 2000. The origins patterns and implications of human spontaneous mutation. Nature Reviews Genetics **1**:40-47.

Dermitzakis, E. T., A. Reymond, and S. E. Antonarakis. 2005. Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nature Reviews Genetics **6**:151-157.

Doll, R. 1996. Cancers weakly related to smoking. Br Med Bull **52**:35-49.

Ebersberger, I. et al. 2002. Genomewide comparison of DNA sequences between humans and chimpanzees. American Journal of Human Genetics **70**:1490-1497.

Elango, N. et al. 2008. Mutations of Different Molecular Origins Exhibit Contrasting Patterns of Regional Substitution Rate Variation. Plos Computational Biology **4**:e1000015.

Ellegren, H., and A. K. Fridolfsson. 1997. Male-driven evolution of DNA sequences in birds. Nature Genetics **17**:182-184.

Ellegren, H., N. G. C. Smith, and M. T. Webster. 2003. Mutation rate variation in the mammalian genome. Current Opinion in Genetics & Development **13**:562-568.

Eory, L., D. L. Halligan, and P. D. Keightley. 2010. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol Biol Evol **27**:177-192.

Eyre-Walker, A., and P. D. Keightley. 1999. High genomic deleterious mutation rates in hominids. Nature **397**:344-347.

Eyre-Walker, A., N. H. Smith, and J. Maynard Smith. 1999. How clonal are human mitochondria? Proc. Roy. Soc. Ser. B. **266**:477-483.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Filipski, J. 1988. Why the Rate of Silent Codon Substitutions in Variable within a Vertebrate Genome. Journal of Theoretical Biology **134**:159-164.

Finkel, T., M. Serrano, and M. A. Blasco. 2007. The common biology of cancer and ageing. Nature **448**:767-774.

Fox, E. J., J. J. Salk, and L. A. Loeb. 2009. Cancer Genome Sequencing-An Interim Analysis. Cancer Research **69**:4948-4950.

Frederico, L. A., T. A. Kunkel, and B. R. Shaw. 1990. A Sensitive Genetic Assay for the Detection of Cytosine Deamination - Determination of Rate Constants and the Activation-Energy. Biochemistry **29**:2532-2537.

Fredman, D. et al. 2004. Complex SNP-related sequence variation in segmental genome duplications. Nat Genet **36**:861-866.

Fryxell, K. J., and W. J. Moon. 2005. CpG mutation rates in the human genome are highly dependent on local GC content (vol 22, pg 650, 2005). Molecular Biology and Evolution **22**:1159-1159.

Fryxell, K. J., and E. Zuckerkandl. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. Molecular Biology and Evolution **17**:1371-1383.

Futreal, P. A. et al. 2004. A census of human cancer genes. Nat Rev Cancer **4**:177-183.

Gaffney, D. J., and P. D. Keightley. 2008. Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. BMC Evol Biol **8**:265.

Gaffney, D. J., and P. D. Keightley. 2005. The scale of mutational variation in the murid genome. Genome Research **15**:1086-1094.

Galtier, N. et al. 2006. Mutation hot spots in mammalian mitochondrial DNA. Genome Res **16**:215-222.

Ge, X. et al. 2005. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics **86**:127-141.

Gentleman, R. C. et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5**:R80.

Gibbs, R. A. et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science **316**:222-234.

Gibbs, R. A. et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428**:493-521.

Goetting-Minesky, M. P., and K. D. Makova. 2006. Mammalian male mutation bias: Impacts of generation time and regional variation in substitution rates. Journal of Molecular Evolution **63**:537-544.

Gojobori, T., W. H. Li, and D. Graur. 1982. Patterns of Nucleotide Substitution in Pseudogenes and Functional Genes. Journal of Molecular Evolution **18**:360-369.

Goodman, M. F. 2002. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. Annu Rev Biochem **71**:17-50.

Green, P. et al. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet **33**:514-517.

Greenman, C. et al. 2007. Patterns of somatic mutation in human cancer genomes. Nature **446**:153-158.

Gupta, S. et al. 2008. Predicting Human Nucleosome Occupancy from Primary Sequence. Plos Computational Biology **4**:-.

Haldane, J. B. S. 1947. The Mutation Rate of the Gene for Haemophilia, and Its Segregation Ratios in Males and Females. Annals of Eugenics **13**:262-271.

Hellmann, I. et al. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. American Journal of Human Genetics **72**:1527-1535.

Hellmann, I. et al. 2005. Why do human diversity levels vary at a megabase scale? Genome Research **15**:1222-1231.

Hernandez, R. D. et al. 2007. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-Content in humans. Molecular Biology and Evolution **24**:2196-2202.

Hess, S. T., J. D. Blake, and R. D. Blake. 1994. Wide Variations in Neighbor-Dependent Substitution Rates. Journal of Molecular Biology **236**:1022-1033.

Hey, J. 2010. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. Mol Biol Evol **27**:921-933.

Holmquist, G. P., and J. Filipski. 1994. Organization of Mutations Along the Genome - a Prime Determinant of Genome Evolution. Trends in Ecology & Evolution **9**:65-69.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18**:337-338.

Hughes, A. L., and M. Nei. 1988. Pattern of Nucleotide Substitution at Major Histocompatibility Complex Class-I Loci Reveals Overdominant Selection. Nature **335**:167-170.

Hwang, D. G., and P. Green. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proceedings of the National Academy of Sciences of the United States of America **101**:13994-14001.

Irizarry, R. A. et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4**:249-264.

Jeffreys, A. J., and C. A. May. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat Genet **36**:151-156.

Johnson, R. E. et al. 2000. Eukaryotic polymerases iota and zeta act sequentially to bypass DNA lesions. Nature **406**:1015-1019.

Keller, I., D. Bensasson, and R. A. Nichols. 2007. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet **3**:e22.

Kennard, O., and W. N. Hunter. 1991. Single-Crystal X-Ray-Diffraction Studies of Oligonucleotides and Oligonucleotide Drug Complexes. Angewandte Chemie-International Edition in English **30**:1254-1277.

Kinlen, L. J., and E. Rogot. 1988. Leukaemia and smoking habits among United States veterans. Bmj **297**:657-659.

Kitts, A., and S. T. Sherry. 2010. NCBI Handbook: The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation.

Kondrashov, A. S. 1995. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? J Theor Biol **175**:583-594.

Kondrashov, A. S. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Human Mutation **21**:12-27.

Kondrashov, F. A., and A. S. Kondrashov. 2010. Measurements of spontaneous rates of mutations in the recent past and the near future. Philosophical Transactions of the Royal Society B-Biological Sciences **365**:1169-1176.

Kong, A. et al. 2002. A high-resolution recombination map of the human genome. Nat Genet **31**:241-247.

Krawczak, M., E. V. Ball, and D. N. Cooper. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. American Journal of Human Genetics **63**:474-488.

Kunkel, T. A. 1985. The Mutational Specificity of DNA Polymerase-Beta during Invitro DNA-Synthesis - Production of Frameshift, Base Substitution, and Deletion Mutations. Journal of Biological Chemistry **260**:5787-5796.

Kunkel, T. A., and A. Soni. 1988. Mutagenesis by transient misalignment. J. Biol. Chem. **263**:14784-14789.

Kuper, H., P. Boffetta, and H. O. Adami. 2002. Tobacco use and cancer causation: association by tumour type. J Intern Med **252**:206-224.

Kurtz, S. et al. 2004. Versatile and open software for comparing large genomes. Genome Biol **5**:R12.

Lander, E. S. et al. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860-921.

Lee, J. B. et al. 2006. DNA primase acts as a molecular brake in DNA replication. Nature **439**:621-624.

Lercher, M. J., and L. D. Hurst. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. Trends in Genetics **18**:337-340.

Lercher, M. J., E. J. Williams, and L. D. Hurst. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. Mol Biol Evol **18**:2032-2039.

Leroy, J. L. et al. 1988. Characterization of Base-Pair Opening in Deoxynucleotide Duplexes Using Catalyzed Exchange of the Imino Proton. Journal of Molecular Biology **200**:223-238.

Ley, T. J. et al. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature **456**:66-72.

Li, W. H., C. I. Wu, and C. C. Luo. 1984. Nonrandomness of Point Mutation as Reflected in Nucleotide Substitutions in Pseudogenes and Its Evolutionary Implications. Journal of Molecular Evolution **21**:58-71.

Li, W. H., S. Yi, and K. Makova. 2002. Male-driven evolution. Curr Opin Genet Dev **12**:650-656.

Loeb, L. A. 1985. Apurinic Sites as Mutagenic Intermediates. Cell **40**:483-484.

Loeb, L. A., J. H. Bielas, and R. A. Beckman. 2008. Cancers exhibit a mutator phenotype: Clinical implications. Cancer Research **68**:3551-3557.

Loeb, L. A., and R. J. Monnat. 2008. DNA polymerases and human disease. Nature Reviews Genetics **9**:594-604.

Lunter, G., and J. Hein. 2004. A nucleotide substitution model with nearest-neighbour interactions. Bioinformatics **20 Suppl 1**:i216-223.

Lynch, M. 2010. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci U S A **107**:961-968.

Makova, K. D., and W. H. Li. 2002. Strong male-driven evolution of DNA sequences in humans and apes. Nature **416**:624-626.

Makova, K. D., S. Yang, and F. Chiaromonte. 2004. Insertions and deletions are male biased too: a whole-genome analysis in rodents. Genome Res **14**:567-573.

Malcom, C. M., G. J. Wyckoff, and B. T. Lahn. 2003. Genic mutation rates in mammals: Local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. Molecular Biology and Evolution **20**:1633-1641.

Malhi, R. S. et al. 2007. MamuSNP: a resource for Rhesus Macaque (Macaca mulatta) genomics. PLoS ONE **2**:e438.

Malyarchuk, B. A., and I. B. Rogozin. 2004. Mutagenesis by transient misalignment in the human mitochondrial DNA control region. Ann Hum Genet **68**:324-339.

Marais, G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends in Genetics **19**:330-338.

Mardis, E. R. 2008. The impact of next-generation sequencing technology on genetics. Trends in Genetics **24**:133-141.

Mardis, E. R. et al. 2009. Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. New England Journal of Medicine **361**:1058-1066.

Matassi, G., P. M. Sharp, and C. Gautier. 1999. Chromosomal location effects on gene sequence evolution in mammals. Current Biology **9**:786-791.

Mcdonald, J. H., and M. Kreitman. 1991. Adaptive Protein Evolution at the Adh Locus in Drosophila. Nature **351**:652-654.

McVean, G. 2000. Evolutionary genetics: What is driving male mutation? Current Biology **10**:R834-R835.

McVean, G. 2009. Discussion on the 1000 genome project.

McVean, G., P. Awadalla, and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160**:1231-1241.

McVean, G. T., and L. D. Hurst. 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. Nature **386**:388-392.

Meuth, M. 1989. The Molecular-Basis of Mutations Induced by Deoxyribonucleoside Triphosphate Pool Imbalances in Mammalian-Cells. Experimental Cell Research **181**:305-316.

Meyer, S., G. Weiss, and A. von Haeseler. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics **152**:1103-1110.

Mikkelsen, T. S. et al. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**:69-87.

Mirkin, E. V., and S. M. Mirkin. 2007. Replication fork stalling at natural impediments. Microbiology and Molecular Biology Reviews **71**:13-35.

Mitchell, A. A. et al. 2004. Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. Bioinformatics **20**:1022-1032.

Miyata, T. et al. 1987. Male-Driven Molecular Evolution - a Model and Nucleotide-Sequence Analysis. Cold Spring Harbor Symposia on Quantitative Biology **52**:863-867.

Musumeci, L. et al. 2010. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. Hum Mutat **31**:67-73.

Nachman, M. W., and S. L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics **156**:297-304.

Nakamura, J., and J. A. Swenberg. 1999. Endogenous apurinic/apyrimidinic sites in genomic DNA of mammalian tissues. Cancer Res **59**:2522-2526.

NIEHS-SNPs. 2008. NIEHS Environmental Genome Project. Univer-
sity of Washington, Seattle:http://egp.gs.washington.edu.

Nielsen, R. et al. 2007. Recent and ongoing selection in the human genome. Nat Rev Genet **8**:857-868.

Ophir, R., and D. Graur. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene **205**:191-202.

Ozsolak, F. et al. 2007. High-throughput mapping of the chromatin structure of human promoters. Nat Biotechnol **25**:244-248.

Paques, F., and J. E. Haber. 1999. Multiple pathways of recombination induced by double-strand breaks in Saccharomyces cerevisiae. Microbiol Mol Biol Rev **63**:349-404.

Pfeifer, G. P. et al. 2002. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene **21**:7435-7451.

Pfeifer, G. P., Y. H. You, and A. Besaratinia. 2005. Mutations induced by ultraviolet light. Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis **571**:19-31.

Pfeiffer, P. 1998. The mutagenic potential of DNA double-strand break repair. Toxicol Lett **96-97**:119-129.

Pink, C. J., and L. D. Hurst. 2010. Timing of Replication Is a Determinant of Neutral Substitution Rates but Does Not Explain Slow Y Chromosome Evolution in Rodents. Molecular Biology and Evolution **27**:1077-1086.

Pink, C. J. et al. 2009. Evidence That Replication-Associated Mutation Alone Does Not Explain Between-Chromosome Differences In Substitution Rates. Genome Biology and Evolution:13-22.

Pleasance, E. D. et al. 2009a. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature **463**:191-196.

Pleasance, E. D. et al. 2009b. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature **463**:184-190.

Prendergast, J. G. D. et al. 2007. Chromatin structure and evolution in the human genome. Bmc Evolutionary Biology **7**:-.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci **13**:235-238.

Rattray, A. J., and J. N. Strathern. 2003. Error-prone DNA polymerases: when making a mistake is the only way to get ahead. Annu Rev Genet **37**:31-66.

Rieder, M. 2010. Personal Communication. University of Washington, Seattle, USA.

Rogozin, I. B., and Y. I. Pavlov. 2003. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. Mutation Research-Reviews in Mutation Research **544**:65-85.

Roth, Y. F. 1987. Eucaryotic primase. Eur J Biochem **165**:473-481.

Russell, L. B. 1999. Significance of the perigametic interval as a major source of spontaneous mutations that result in mosaics. Environmental and Molecular Mutagenesis **34**:16-23.

Sachidanandam, R. et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409**:928-933.

SantaLucia, J., H. T. Allawi, and A. Seneviratne. 1996. Improved nearest-neighbor parameters for predicting DNA duplex stability. Biochemistry **35**:3555-3562.

SeattleSNPs. 2008. NHLBI Program for Genomic Applications. SeattleSNPs, Seattle:http://pga.gs.washington.edu.

Sherry, S. T. et al. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res **29**:308-311.

Shiratori, A. et al. 1995. Assignment of the 49-kDa (PRIM1) and 58-kDa (PRIM2A and PRIM2B) subunit genes of the human DNA primase to chromosome bands 1q44 and 6p11.1-p12. Genomics **28**:350-353.

Silva, J. C., and A. S. Kondrashov. 2002. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. Trends in Genetics **18**:544-547.

Skaletsky, H. et al. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature **423**:825-837.

Slater, G. S., and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6**:31.

Smith, N. G. C., M. T. Webster, and H. Ellegern. 2003. A low rate of simultaneous double-nucleotide mutations in primates. Molecular Biology and Evolution **20**:47-53.

Smith, N. G. C., M. T. Webster, and H. Ellegren. 2002. Deterministic mutation rate variation in the human genome. Genome Research **12**:1350-1356.

Stagnaro, E. et al. 2001. Smoking and hematolymphopoietic malignancies. Cancer Causes Control **12**:325-334.

Stamatoyannopoulos, J. A. et al. 2009. Human mutation rate associated with DNA replication timing. Nature Genetics **41**:393-395.

Stenson, P. D. et al. 2009. The Human Gene Mutation Database: 2008 update. Genome Med **1**:13.

Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics **68**:978-989.

Stoneking, M. 2000. Hypervariable sites in the mtDNA control region are mutational hotspots. Am. J. Hum. Genet. **67**:1029-1032.

Su, A. I. et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A **101**:6062-6067.

Sueoka, N. 1992. Directional Mutation Pressure, Selective Constraints, and Genetic Equilibria. Journal of Molecular Evolution **34**:95-114.

Takai, D., and P. A. Jones. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A **99**:3740-3745.

Taylor, F. et al. 2006. Strong and weak male mutation bias at different sites in the primate genomes: Insights from the human-chimpanzee comparison. Molecular Biology and Evolution **23**:565-573.

Templeton, A. R. et al. 2000. Recombinational and mutational hotspots within the human lipoprotein lipase gene. Am J Hum Genet **66**:69-83.

Tian, D. C. et al. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature **455**:105-U170.

Todorova, A., and G. A. Danieli. 1997. Large majority of single-nucleotide mutations along the dystrophin gene can be explained by more than one mechanism of mutagenesis. Human Mutation **9**:537-547.

Topal, M. D., and J. R. Fresco. 1976. Complementary Base-Pairing and Origin of Substitution Mutations. Nature **263**:285-289.

Underhill, P. A. et al. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Annals of Human Genetics **65**:43-62.

Venter, J. C. et al. 2001. The sequence of the human genome. Science **291**:1304-1351.

Vijg, J. 2000. Somatic mutations and aging: a re-evaluation. Mutat Res **447**:117-135.

Wakabayashi, I. et al. 1994. A case-control study on risk factors for leukemia in a district of Japan. Intern Med **33**:198-203.

Wall, J. D. 2003. Estimating ancestral population sizes and divergence times. Genetics **163**:395-404.

Walser, J. C., L. Ponger, and A. V. Furano. 2008. CpG dinucleotides and the mutation rate of non-CpG DNA. Genome Research **18**:1403-1414.

Waterston, R. H. et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.

Watterson, G. A. 1975. Number of Segregating Sites in Genetic Models without Recombination. Theoretical Population Biology **7**:256-276.

Williams, E. J. B., and L. D. Hurst. 2002. Clustering of tissue-specific genes underlies much of the similarity in rates of protein evolution of linked genes. Journal of Molecular Evolution **54**:511-518.

Williams, E. J. B., and L. D. Hurst. 2000. The proteins of linked genes evolve at similar rates. Nature **407**:900-903.

Wilson, D. M., 3rd, and D. Barsky. 2001. The major human abasic endonuclease: formation, consequences and repair of abasic lesions in DNA. Mutat Res **485**:283-307.

Wilson, D. M., 3rd et al. 1995. Incision activity of human apurinic endonuclease (Ape) at abasic site analogs in DNA. J Biol Chem **270**:16002-16007.

Winckler, W. et al. 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. Science **308**:107-111.

Wolfe, K. H., and P. M. Sharp. 1993. Mammalian Gene Evolution - Nucleotide-Sequence Divergence between Mouse and Rat. Journal of Molecular Evolution **37**:441-456.

Wolfe, K. H., P. M. Sharp, and W. H. Li. 1989. Mutation-Rates Differ among Regions of the Mammalian Genome. Nature **337**:283-285.

Xue, Y. L. et al. 2009. Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree. Current Biology **19**:1453-1457.

Ying, H. et al. 2010. Evidence that Localized Variation in Primate Sequence Divergence Arises from an Influence of Nucleosome Placement on DNA Repair. Molecular Biology and Evolution **27**:637-649.

Yu, N. et al. 2001. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol Biol Evol **18**:214-222.

Zhang, W. et al. 2007. Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. Journal of Molecular Evolution **65**:207-214.

Zhao, Z., and E. Boerwinkle. 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: A study of 2.6 million polymorphisms across the human genome. Genome Research **12**:1679-1686.

Zhao, Z. et al. 2003. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene **312**:207-213.

Appendices

Appendix 2.1. A description of the two-state model for quantifying cryptic variation in the mutation rate.

Let us assume that a proportion, α, of sites are hypermutable and that they mutate at a rate β, whereas normal sites mutate at a rate of β^*. Let us assume that hypermutable sites are always hypermutable. The average rate of mutation in the sequence is therefore

$$\bar{\mu} = \alpha\beta + (1-\alpha)\beta^* \qquad \text{(A2.1)}$$

If we arbitrarily set the average rate of mutation to one then we can express β^* in terms of α and β.

$$\beta^* = \frac{1-\alpha\beta}{1-\alpha} \qquad \text{(A2.2)}$$

Let the average probability of detecting a SNP in humans and chimpanzees be μ_h and μ_c respectively, then the expected number of coincident SNPs is

$$P = \alpha\mu_h\mu_c\beta^2 + (1-\alpha)\mu_h\mu_c\beta^{*2} \qquad \text{(A2.3)}$$

If there is no variation in the mutation rate then the expected number of coincident SNPs is

$$P_0 = \mu_h\mu_c \qquad \text{(A2.4)}$$

So the observed number of SNPs over the number expected with no variation is

$$Z = \frac{P}{P_0} = \alpha\beta^2 + (1-\alpha)\beta^{*2} \qquad \text{(A2.5)}$$

The expected excess of coincident SNPs therefore depends upon the proportion of sites that are hypermutable and the ratio β/β^*. The observed excess of coincident SNPs at non-CpG sites is 1.98; we therefore set α to a range of values and solved equation A2.5 for β/β^*. The results are shown in appendix 2.4.

Appendix 2.2

a)

	C/T	G/A	C/A	G/T	C/G	A/T
C/T	1120	1	103	46	101	53
G/A	3	1105	53	99	77	60
C/A	117	54	294	0	46	23
G/T	57	138	0	308	35	13
C/G	79	89	34	30	167	1
A/T	57	56	28	20	0	561

b)

	C/T	G/A	C/A	G/T	C/G	A/T
C/T	2.23		0.99	1.12	1.23	1.04
G/A		2.17	1.30	0.96	0.91	1.20
C/A	1.13	1.15	5.12		1.51	1.30
G/T	1.25	1.34		5.38	1.14	0.71
C/G	1.05	1.17	1.23	1.09	2.78	
A/T	1.07	1.07	1.69	1.16		14.90

Appendix 2.2. The pattern of coincident SNPs for non-CpG sites. a) Gives the number of times a particular SNP in humans is found opposite a particular SNP in chimpanzees. b) The observed number of SNPs over the number expected with simple context effects. Note, some of the observed values are greater than when we included CpG dinucleotides. This is because we re-ran the analysis and when a chimp SNP had matched multiple human sequences we chose a sequence in which the human SNP was not involved in a CpG. The tables show that there is an excess of coincident SNPs above that expected given simple context effects, particularly along the leading diagonal, for non-CpG sites.

Appendix 2.3

	All sites			*Non-CpG sites*		
GC content	*Observed*	*Expected*	*Ratio*	*Observed*	*Expected*	*Ratio*
Lower quartile	*2793*	*1420*	*1.97 (0.04)*	*1633*	*752*	*2.17 (0.05)*
Upper quartile	*2890*	*1862*	*1.55 (0.03)*	*606*	*367*	*1.65 (0.07)*

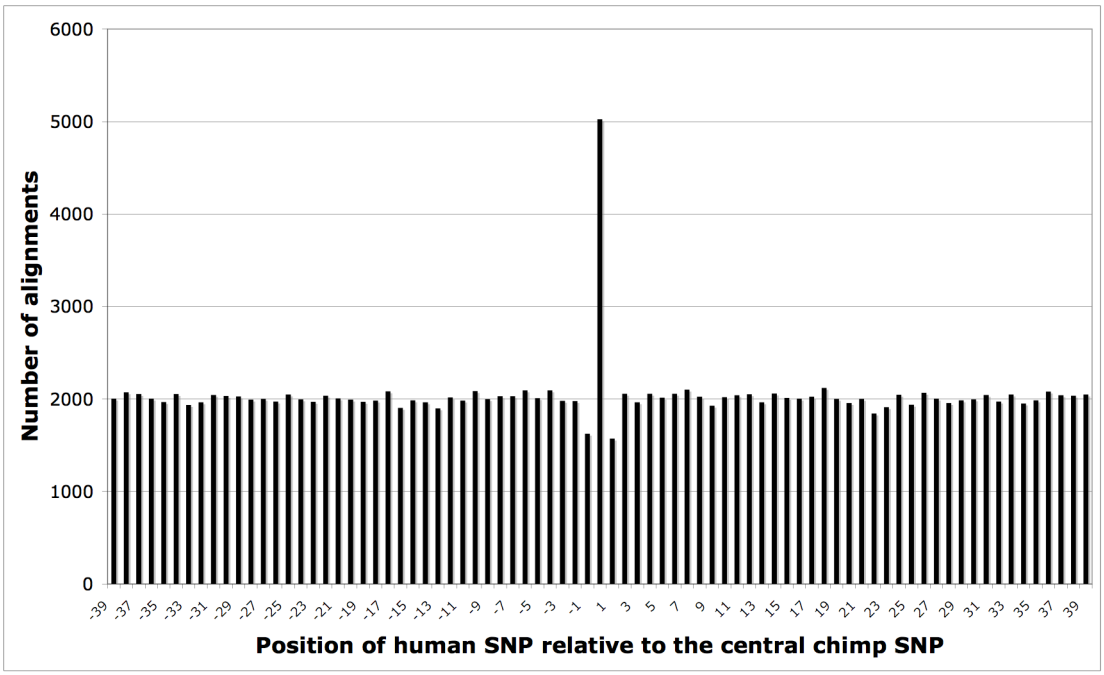
Appendix 2.3. The observed and expected numbers of coincident SNPs in the alignments with high or low GC content. Standard errors are given for the ratio. There is a significant excess of coincident SNPs in both the upper and lower GC quartile datasets.

Appendix 2.4

α	β/β^*
0.001	33
0.002	24
0.01	12
0.02	9.2
0.1	5.9
0.2	5.9

Appendix 2.4. The results from the simple two-rate model presented in appendix 2.1. The results are similar to those obtained from the log-normal model detailed in the main text and consequently we focus solely on that model.

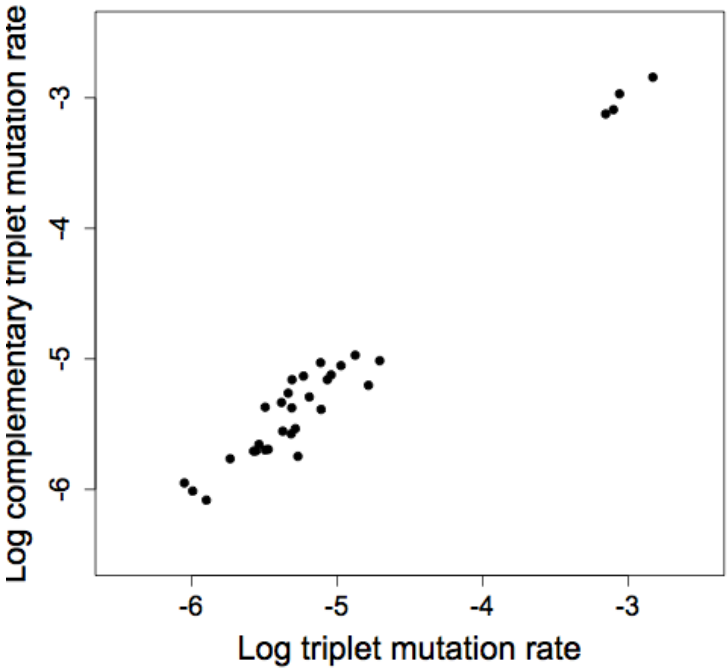
Appendix 2.5



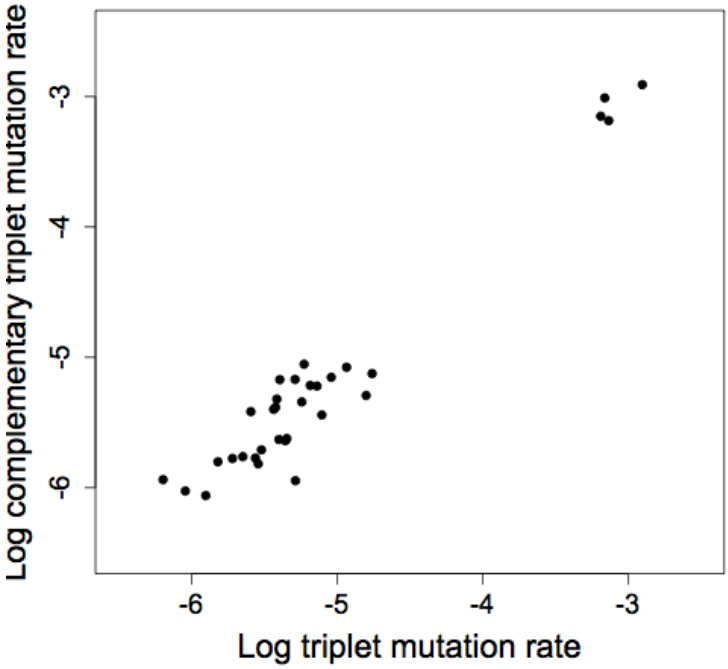
Appendix 2.5. The number of human SNPs at each site of the human-chimpanzee alignments used in the analysis excluding CpG sites. The slight deficit of human SNPs adjacent to the chimpanzee is due to the fact that the adjacent sites are more likely to be inferred to be within a CpG because the chimp SNP might contain either C or G. For example, if the human SNP at +1 is G/A and the chimp SNP is C/G this would be called a potential CpG site and excluded. The graph shows an excess of coincident SNPs at position zero, with the frequency of human SNPs at all other positions being relatively uniform, showing that there is no tendency for single SNPs to cluster.

Appendix 2.6

a)

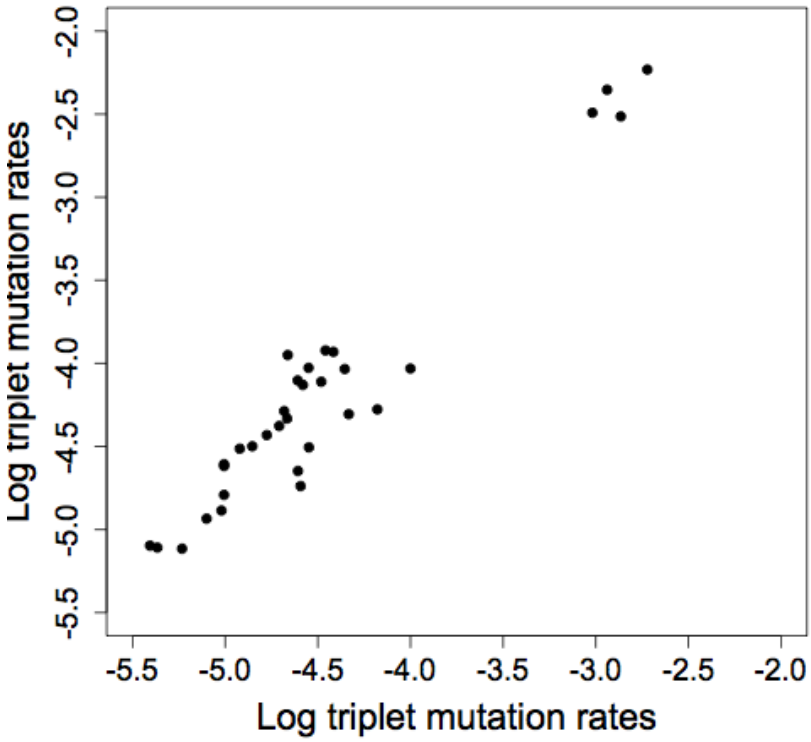


b)



Appendix 2.6. The rate of mutation for each triplet on the coding strand (delete) plotted against its reverse complement for (a) all genes and (b) genes expressed in the testes. The graphs show that the relative rates of mutation are similar on the coding and non-coding strand in all genes and those expressed in the testes, and consequently the excess of coincident SNPs is not a result of strand asymmetry in the pattern of mutation.

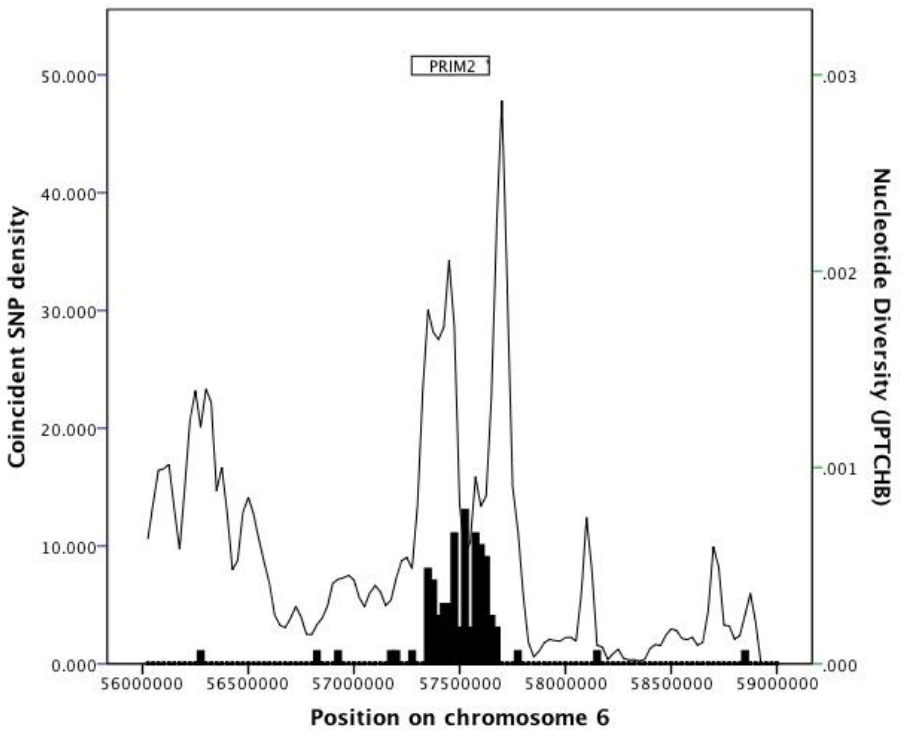
Appendix 2.7



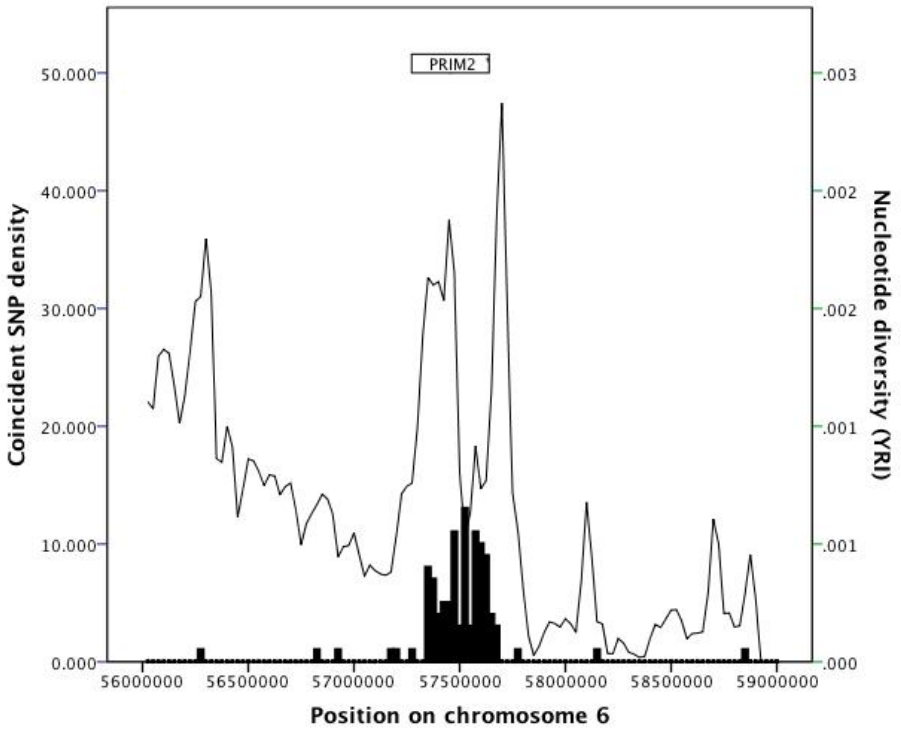
Apendix 2.7. The rate of mutation for each triplet in the GC-rich alignments (x-axis) versus the rate of mutation in the GC-poor alignments (y-axis). The graph shows that the relative rates of mutation are similar in GC-rich and GC-poor alignments, and thus the excess of coincident SNPs is not a consequence of mis-inferring triplet mutation rates in different contexts.

Appendix 3.1

a)



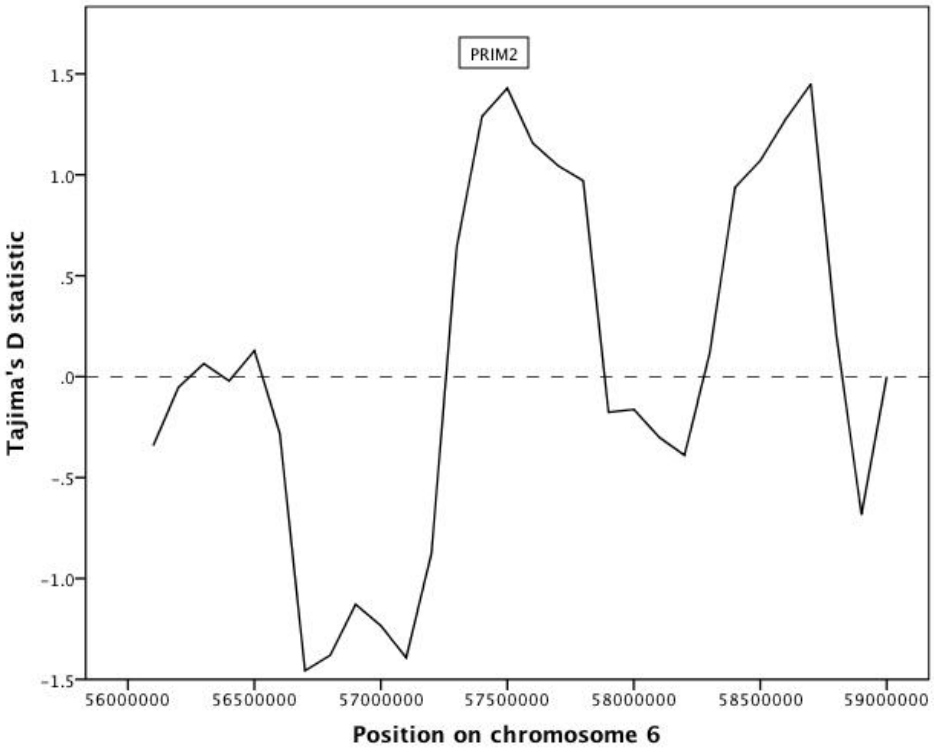
b)



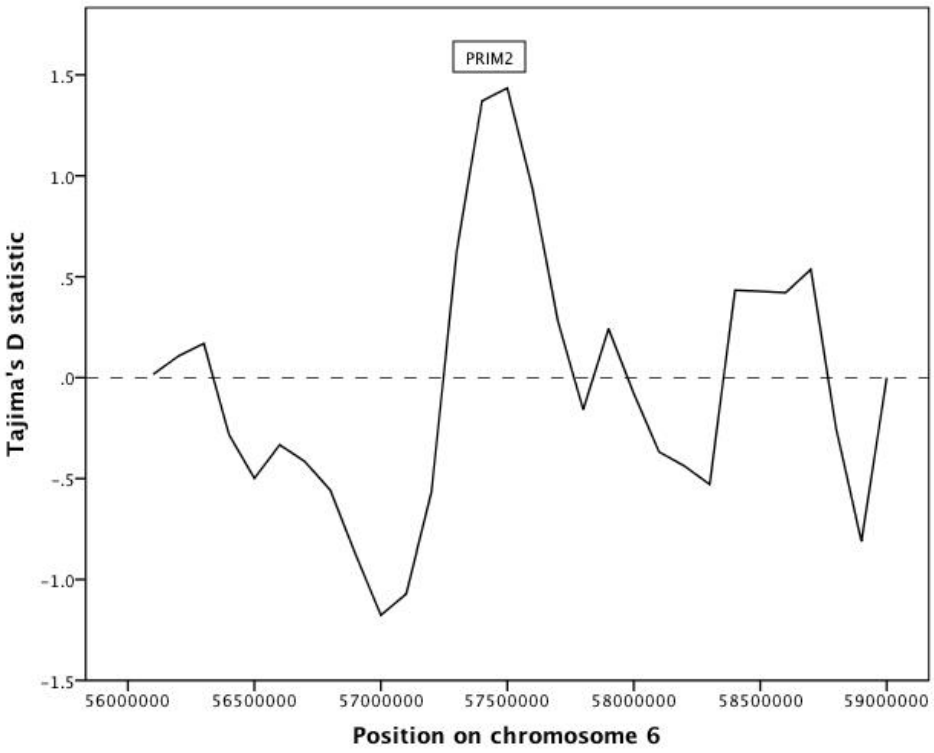
Appendix 3.1: The nucleotide diversity across the region containing PRIM2 for (a) the CHB+JPT and (b) the YRI populations. The figure shows a sliding window of nucleotide diversity every 25kb, with window size of 50kb as a line graph corresponding to the right hand axis, with the coincident SNP densities as a bar chart corresponding to the left hand axis. The region with the highest density of coincident SNPs also has a relatively high single SNP density in most cases.

Appendix 3.2

a)



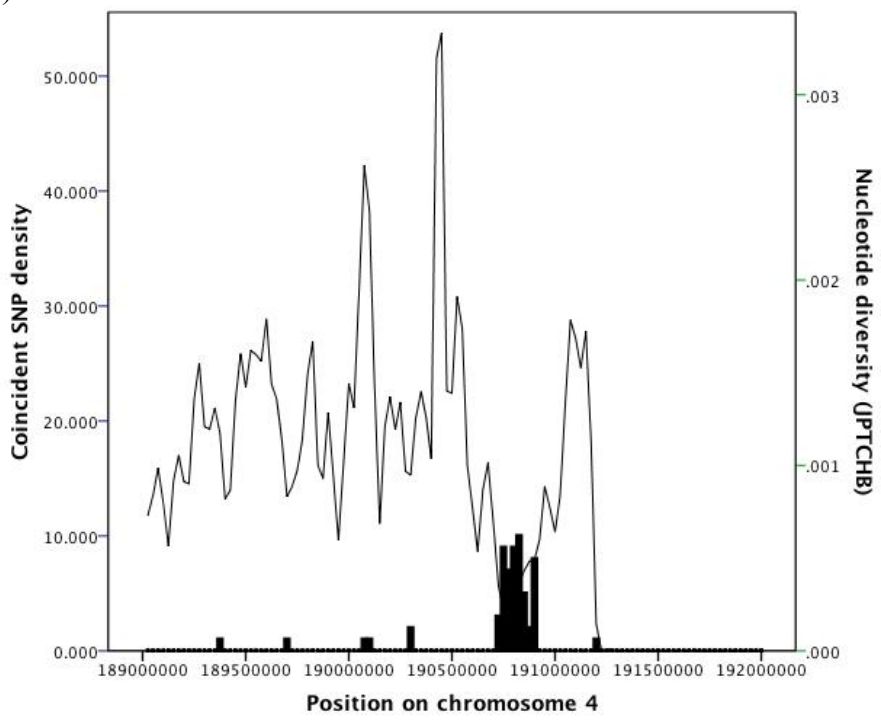
b)



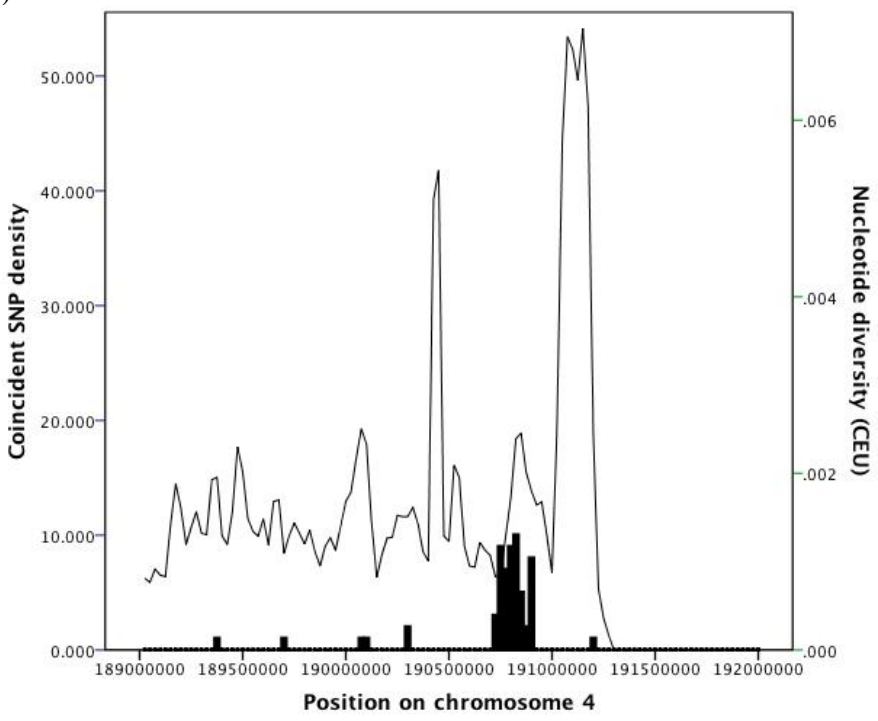
Appendix 3.2: The Tajima's D statistic across the region containing PRIM2 for (a) the CHB+JPT and (b) the YRI populations. The figure shows a sliding window of nucleotide diversity every 100kb, with window size of 200kb. The region containing PRIM2 has a higher Tajima's D statistic than most surrounding regions, and thus PRIM2 may be under balancing selection.

Appendix 3.3

a)



b)



Appendix 3.3: The nucleotide diversity across the region on chromosome 4 for (a) the CHB+JPT and (b) the CEU populations. The figure shows a sliding window of nucleotide diversity every 25kb, with window size of 50kb as a line graph corresponding to the right hand axis, with the coincident SNP densities as a bar chart corresponding to the left hand axis. The region on chromosome 4 with a high concentration of coincident SNPs has one of the lowest densities of single SNPs in the region

Appendix 4.1

Appendix 4.1: Tri-allelic SNP locations and allele frequencies. Genomic locations are from Ensembl release 56 (www.ensembl.org/Homo_sapiens).

Gene	Chr	Genomic Location	Major Allele	Freq.	Minor allele	Freq.	Minor allele	Freq.
ABCB1	7	87134535	G	170	C	14	A	2
ABCB4	7	87081432	T	164	G	4	A	2
ABCC1	16	16169574	G	167	C	10	A	1
ADH5	4	99994622	G	161	C	14	T	1
ALDH1A2	15	58249834	C	168	A	21	T	1
ALOX5AP	13	31331915	G	75	A	11	C	8
APOH	17	64220092	C	70	G	18	A	2
APP	21	27503569	C	162	T	17	A	7
APP	21	27270116	G	162	C	1	A	1
ATRX	X	76842900	G	173	A	3	T	2
BNIP3	10	133785129	T	116	G	53	C	1
BRCA1	17	41218426	C	174	G	1	T	1
CASP9	1	15844131	C	178	G	1	T	1
CCND1	11	69462202	C	167	T	3	A	2
CCNI	4	77974190	C	164	G	13	A	3
CD247	1	167412070	C	173	T	8	A	1
CD27	12	6557735	T	176	G	3	C	1
CHUK	10	101961390	A	87	T	5	G	2
CP	3	148894357	A	184	C	5	G	1
CTNND2	5	11822213	C	188	T	1	A	1
CTNND2	5	11496373	G	122	C	65	A	3
CYP2C8	10	96797158	G	174	A	5	T	1
CYP4F2	19	15996483	C	67	T	3	G	1
CYP4V2	4	187115996	T	126	G	63	A	1
DCN	12	91542527	A	79	G	14	T	1
DCN	12	91540103	A	74	G	15	T	3
DDB1	11	61080839	T	183	A	2	C	1
DNAJC3	13	96442178	T	167	C	11	G	2
ECE1	1	21604667	A	185	G	2	T	1
EIF2AK2	2	37353832	C	169	A	6	T	1
ERCC3	2	128042109	G	178	T	1	A	1
ERCC4	16	14025590	C	138	G	22	A	4
ERCC8	5	60195619	G	161	A	15	T	4
F9	X	138613499	G	91	A	2	T	1
F9	X	138616642	C	49	A	43	G	2
FANCA	16	89875207	G	174	T	1	A	1
FANCA	16	89814818	C	91	A	65	T	2
FANCD2	3	10105399	C	173	T	10	A	1
FGF1	5	141980820	C	136	A	33	G	1
FGF20	8	16858885	G	104	A	50	T	26
FGF20	8	16858876	A	177	C	2	T	1

Appendix 4.1: Tri-allelic SNP locations and allele frequencies. Genomic locations are from Ensembl release 56 (www.ensembl.org/Homo_sapiens). Table continued from previous page.

Gene	Chr	Genomic Location	Major Allele	Freq.	Minor allele	Freq.	Minor allele	Freq.
FGF5	4	81193146	G	171	A	6	T	1
FGF5	4	81197802	C	183	T	2	A	1
FGFR1	8	38292902	A	163	C	4	G	1
GAD2	10	26518418	T	128	C	39	A	9
GAS6	13	114542957	G	69	C	16	A	9
GCLC	6	53373276	G	167	C	2	A	1
GPX7	1	53070740	A	184	G	3	C	1
GSR	8	30566786	C	162	T	11	A	7
GSTA4	6	52843758	A	160	T	24	C	4
HPGD	4	175423755	A	57	T	21	C	4
IGFBP7	4	57904663	C	115	T	37	G	10
IL4R	16	27356680	G	89	A	2	T	1
LIPE	19	42930288	T	79	A	4	C	1
MAPK1	22	22139890	A	115	G	32	C	1
MAPK9	5	179696997	T	141	C	39	G	8
MAPT	17	44065708	A	172	C	7	T	1
MB	22	36007803	G	182	T	5	A	3
MCM6	2	136628183	C	164	A	10	G	2
MDM2	12	69214894	C	164	T	14	G	2
MMP12	11	102735103	C	139	T	45	A	2
MMP16	8	89217739	G	173	A	12	T	1
MNAT1	14	61241111	G	163	A	14	C	3
MSH6	2	48032937	C	135	T	40	G	1
MUC5B	11	1161713	G	96	C	93	A	1
MUC5B	11	1280099	T	187	C	2	A	1
MYBPC3	11	47365372	G	172	T	1	A	1
NBN	8	90950688	G	177	C	2	T	1
ORC2L	2	201802566	G	174	C	1	A	1
OXSR1	3	38209896	G	177	A	7	T	6
PARP2	14	20822219	G	152	T	14	C	2
PCSK9	1	55507314	G	85	T	8	A	1
PIK3R5	17	8789532	G	87	C	4	A	1
PKM2	15	72506120	G	154	T	1	A	1
PLA2G4A	1	186896603	A	106	G	67	C	1
PLA2G6	22	38521077	G	160	A	7	T	1
PMS1	2	190690517	A	135	C	10	G	3
PNKP	19	50368116	G	139	T	32	C	1
PNKP	19	50366164	G	129	T	42	A	5
PON1	7	94943123	A	86	C	7	G	1
PON3	7	94992644	T	85	C	8	A	1
PPARG	3	12434070	G	76	C	4	A	4

Appendix 4.1: Tri-allelic SNP locations and allele frequencies. Genomic locations are from Ensembl release 56 (www.ensembl.org/Homo_sapiens). Table continued from previous page.

Gene	Chr	Genomic Location	Major Allele	Freq.	Minor allele	Freq.	Minor allele	Freq.
PRDX3	10	120937828	C	168	A	5	T	1
PRKCB	16	24056018	C	125	A	64	T	1
PRKDC	8	48801948	G	171	C	2	T	1
PRKDC	8	48791668	G	176	C	1	T	1
PSD4	2	113945902	C	53	T	38	A	3
PSD4	2	113947847	G	88	T	5	A	1
PSD4	2	113951924	C	62	A	25	T	7
PTCH2	2	45303959	C	157	T	4	G	1
RAD17	5	68694169	G	88	A	85	C	1
RB1	13	48947469	T	166	G	7	A	1
REV3L	6	111626944	G	176	T	1	A	1
RIPK1	6	3104135	T	84	G	5	C	3
SCARA3	8	27509262	G	184	A	4	C	2
SLC6A3	5	1400241	C	172	T	2	A	2
SNCA	4	90673770	C	116	G	49	T	21
STAT4	2	191898949	G	62	A	30	T	2
SULT1E1	4	70718924	C	165	T	8	A	1
SULT2A1	19	48374950	G	175	T	12	A	1
TDP1	14	90499324	C	165	G	24	A	1
TGFBR2	3	30674339	G	147	A	2	T	1
TGM2	20	36792842	T	109	A	71	C	2
TNFRSF8	1	12170425	G	166	T	12	A	2
TNFRSF9	1	7987558	G	164	T	1	A	1
TRIM5	11	5699801	T	141	G	24	A	1
TUBA3C	13	19752039	C	182	A	5	T	1
UGT2B4	4	70347172	T	145	G	23	A	2
UHRF1	19	4928699	G	102	C	77	A	1
VLDLR	9	2629029	A	66	C	16	G	8
WRN	8	31023686	G	149	A	30	T	1
XPA	9	100438652	T	149	G	26	C	1
XRCC1	19	44053360	T	164	C	3	A	3

Appendix 5.1

Appendix 5.1: Outlier regions for the frequency of cancer mutations per MB. * indicates those genes associated with at least one form of cancer in the OMIM database (http://www.ncbi.nlm.nih.gov/omim).

Cancer	Chromosome	Region (MB)	Number of mutations	Genes
SCLC	X	61-62	77	None
SCLC	6	57-58	58	ZNF451 BAG2 RAB23 PRIM2
SCLC	4	136-137	40	None
SCLC	3	95-96	39	None
SCLC	8	138-139	39	None
SCLC	12	128-129	36	TMEM132C
SCLC	2	79-80	36	SNAR-H
SCLC	8	112-113	35	None
SCLC	14	41-42	35	None
SCLC	5	24-25	35	CDH10
SCLC	8	88-89	34	CNBD1 DCAF4L2
SCLC	14	42-43	34	LRFN5
SCLC	2	81-82	33	None
SCLC	8	111-112	33	None
SCLC	4	135-136	33	PABPC4L
SCLC	1	239-240	33	CHRM3
Skin cancer	7	119-120	50	KCND2
Skin cancer	13	93-94	47	GPC5* GPC6
Skin cancer	7	57-58	47	ZNF479 LOC642006 ZNF716
Skin cancer	7	53-54	46	POM121L12
Skin cancer	3	68-69	44	FAM19A1 FAM19A4
Skin cancer	9	120-121	42	ASTN2 TLR4*
Skin cancer	12	61-62	41	None
Skin cancer	3	162-163	41	None
Skin cancer	8	47-48	41	BEYLA

Appendix 5.1: Outlier regions for the frequency of cancer mutations per MB. * indicates those genes associated with at least one form of cancer in the OMIM database (http://www.ncbi.nlm.nih.gov/omim). Table continued from previous page.

Cancer	Chromosome	Region (MB)	Number of mutations	Genes
AML1	6	32-33	97	C4A C4B CYP21A2 TNXB ATF6B FKBPL PRRT1 PPT2 EGFL8 AGPAT1 RNF5 RNF5P1 AGER* PBX2 GPSM3 NOTCH4 C6orf10 BTNL2 HLA-DRA* HLA-DRB5 HLA-DRB6 HLA-DRB1 HLA-DQA1 HLA-DQB1 HLA-DQA2 HLA-DQB2 HLA-DOB TAP2 PSMB8 TAP1* PSMB9 PPP1R2P1 HLA-DMB HLA-DMA BRD2 HLA-DOA

Appendix 5.1: Outlier regions for the frequency of cancer mutations per MB. * indicates those genes associated with at least one form of cancer in the OMIM database (http://www.ncbi.nlm.nih.gov/omim). Table continued from previous page.

Cancer	Chromosome	Region (MB)	Number of mutations	Genes
AML1	6	29-30	84	OR2W1 OR2B3 OR2J3 OR2J2 OR14J1 OR5V1 OR12D3 OR12D2 OR11A1 OR10C1 OR2H1 MAS1L* UBD SNORD32B OR2H2 GABBR1 MOG ZFP57 HLA-F LOC285830 IFITM4P HCG4 HLA-G* HLA-H HCG2P7 HCG4P6 HLA-A HCG9 NCRNA00171 HLA-J
AML1	3	75-76	67	FAM86D MIR1324 FRG2C ZNF717

Appendix 5.1: Outlier regions for the frequency of cancer mutations per MB. * indicates those genes associated with at least one form of cancer in the OMIM database (http://www.ncbi.nlm.nih.gov/omim). Table continued from previous page.

Cancer	Chromosome	Region (MB)	Number of mutations	Genes
AML1	12	11-12	50	PRR4* PRH1 TAS2R13 PRH2 TAS2R14 TAS2R50 TAS2R20 TAS2R19 TAS2R31 TAS2R46 TAS2R43 TAS2R30 TAS2R42 PRB3 PRB4 PRB1 PRB2 ETV6*
AML1	2	133-134	48	NCRNA00164 MIR663B GPR39 LYPD1* NCKAP5
AML1	2	37-38	34	VIT STRN HEATR5B CCDC75 EIF2AK2 SULT6B1 CEBPZ C2orf56 PRKD3 QPCT CDC42EP3

Appendix 5.1: Outlier regions for the frequency of cancer mutations per MB. * indicates those genes associated with at least one form of cancer in the OMIM database (http://www.ncbi.nlm.nih.gov/omim). Table continued from previous page.

Cancer	Chromosome	Region (MB)	Number of mutations	Genes
AML2	3	75-76	473	See above
AML2	2	133-134	252	See above
AML2	13	63-64	149	None
AML2	6	32-33	106	See above
AML2	2	132-133	92	POTEE LOC440910 LOC150786 LOC401010 TUBA3D FAM128A LOC150776 CCDC74A C2orf27A C2orf27B NCRNA00164
AML2	2	37-38	90	See above
AML2	4	3-4	89	GRK4 HTT C4orf44 RGS12 HGFAC DOK7 LRPAP1 ADRA2C LOC348926
AML2	9	69-70	81	PGM5P2 LOC440896 FOXD4L6 CBWD6 ANKRD20A4 LOC100133920
AML2	10	127-128	63	LOC100169752 C10orf122 C10orf137 MMP21* UROS BCCIP* DHX32 FANK1 ADAM12

Appendix 5.1: Outlier regions for the frequency of cancer mutations per MB. * indicates those genes associated with at least one form of cancer in the OMIM database (http://www.ncbi.nlm.nih.gov/omim). Table continued from previous page.

Cancer	Chromosome	Region (MB)	Number of mutations	Genes
AML2	12	11-12	62	See above
AML2	1	149-150	59	LOC388692 FCGR1C HIST2H2BF PPIAL4A PPIAL4C PPIAL4B LOC728855 FCGR1A HIST2H3D HIST2H4A* HIST2H4B HIST2H3C HIST2H3A HIST2H2AA3 HIST2H2AA4 HIST2H3A HIST2H3C HIST2H2BE HIST2H2AC HIST2H2AB BOLA1 SV2A SF3B4 MTMR11 OTUD7B